Statistical Analysis of Karcher Means for Random Restricted PSD Matrices

Hengchao Chen
Department of Statistical Sciences
University of Toronto

Xiang Li
School of Mathematical Sciences
Peking University

Qiang Sun
Department of Statistical Sciences
University of Toronto

Abstract

Non-asymptotic statistical analysis is often missing for modern geometry-aware machine learning algorithms due to the possibly intricate nonlinear manifold structure. This paper studies an intrinsic mean model on the manifold of restricted positive semi-definite matrices and provides a non-asymptotic statistical analysis of the Karcher mean. We also consider a general extrinsic signal-plus-noise model, under which a deterministic error bound of the Karcher mean is provided. As an application, we show that the distributed principal component analysis algorithm, LRC-dPCA, achieves the same performance as the full sample PCA algorithm. Numerical experiments lend strong support to our theories.

1 Introduction

Positive semi-definite (PSD) matrices arise in a wide range of applications, such as covariance matrices in statistics (Wainwright, 2019), kernel matrices in machine learning (Hastie et al., 2009), diffusion tensor images in medical imaging (Dryden et al., 2009), semi-definite programming (Journée et al., 2010), and covariance descriptors in image set classification (Wang et al., 2012), just to name a few. From athe geometric perspectiveviewpoint, the cone of PSD matrices is not a vector space, since linear combinations of multiple PSD matrices are not necessarily PSD matrices. Instead, the set of (restricted) PSD matrices of fixed rank has been endowed with different metrics such that it forms a Riemannian manifold (Bonnabel and Sepulchre, 2010; Vandereycken et al., 2013; Massart and Absil, 2020; Neuman et al., 2021). By utilizing the geometric structures, researchers have developed many powerful statistical or computational methods (Faraki et al., 2016; Cornea et al., 2017; Patrangenaru and Ellingson, 2016).

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

One important concept in Riemannian geometry or more generally metric spaces is the Karcher mean (Karcher, 1977). The Karcher mean is often referred to as the Fréchet mean or the barycenter of mass. Given M points $\{\boldsymbol{z}_m\}_{m=1}^M$ on a metric space $(\mathcal{M}, d)$ with distance function $d(\cdot,\cdot)$, the Karcher mean $\widetilde{\boldsymbol{z}}$ of these points is given by

$$\widetilde{\boldsymbol{z}} = \operatorname*{argmin}_{\boldsymbol{z}\in\mathcal{M}} \sum_m d^2(\boldsymbol{z}, \boldsymbol{z}_m). \tag{1.1}$$

When the underlying space is Euclidean, the Karcher mean is reducesd to the arithmetic mean. In general, the existence and computation of the Karcher mean is already complicated due to the possibly intricate non-Euclidean structure (Karcher, 1977; Bini and Iannazzo, 2013). As a result, most works focus on the computation and applications of the Karcher mean, withhile few works providinge statistical guarantees. Statistically, Bhattacharya and Patrangenaru (2003, 2005) establish a large sample theory of the Karcher mean on manifolds with applications to spheres and projective spaces. Bigot and Gendre (2013) shows the minimax optimality of the Karcher mean of discretely sampled curves. In this paper, we consider the manifold of restricted PSD matrices as in (Neuman et al., 2021). In particular, we first study an intrinsic mean model, inspired bywhich is aligned with the geometric structure of the restricted PSD manifold. A non-asymptotic statistical analysis of the Karcher mean is provided under thissuch intrinsic model. We furthAlso, wer consider a general extrinsic signal-plus-noise model, which does not necessarily coincide with the manifold geometry by Neuman et al. (2021). For thissuch general model, we give a deterministic error bound ofor the Karcher mean, which is then used to provide an error bound for and apply the results to a distributed principal component analysis algorithm.

The Karcher mean is closely related to distributed learning problems, especially the divide-and-conquer (DC) framework (Mackey et al., 2011). In distributed learning problems, massive datasets are scattered across distant servers and directly fusing these datasets is extremely challenging due to concerns on communication cost, privacy, data security, and ownership, among others. A commonly used distributed framework is tTo address these problems, the DC framework which first computes local estimators lo-

cally and then aggregate them on the central server, where the last step is often equivalent to computing the Karcher mean on certain manifolds. For example, the divide-and-conquer principal component analysis (PCA) algorithms (Fan et al., 2019; Bhaskara and Wijewardena, 2019; Neuman et al., 2021) essentially compute the Karcher means on the Grasssmann manifold, Euclidean space, or the manifold of restricted PSD matrices, respectively. Motivated by this observation, we give theoretical guarantees of the DC PCA algorithm, LRC-dPCA, proposed in Neuman et al. (2021) by applying our non-asymptotic statistical analysis of the Karcher mean on the restricted PSD manifold. Specifically, we show that given sufficiently large local sample size, LRC-dPCA achieves the same performance as the full sample PCA algorithm, which outputs the top eigenvectors of the covariance matrix based on full data.

Our contributions are three-fold. First, we provide a non-asymptotic statistical analysis of the Karcher mean on the restricted PSD manifold under an intrinsic model. Second, for a generic signal-plus-noise model, we give a deterministic characterization of the Karcher mean and then obtain a deterministic error bound. Third, as an application, we show that LRC-dPCA and full sample PCA share the same performance given sufficiently large local sample size. Numerical experiments are carried out to support our theories.

The rest of this paper proceeds as follows. We conclude this section with a discussion on related works. Section 2 reviews the geometry for restricted PSD matrices proposed in Neuman et al. (2021). Then in Section 3, we provide the theoretical analysis of the Karcher mean on the restricted PSD manifolds. Applications to distributed PCA algorithms are given in Section 4. Numerical experiments are carried out in Section 5 and we give concluding remarks in Section 6. Proofs are left to the Appendix.

1.1 Related work

Manifolds of PSD matrices The cone of symmetric positive definite (SPD) matrices is not a vector space. It can be viewed as different Riemannian manifolds when endowed with different metrics, such as the affine-invariant metric (Moakher, 2005) and the Log-Euclidean metric (Arsigny et al., 2007). It is, however, non-trivial to generalize these metrics to the rank-deficient (PSD) case. To this end, Bonnabel and Sepulchre (2010) treated a PSD matrix of rank K in a quotient space as a K-dimensional subspace coupled with a K-by-K SPD matrix and then endowed the manifold of PSD matrices with a weighted product metric. Using this geometry, Bonnabel et al. (2013) developed a rank-preserving geometric mean of PSD matrices. Later, Vandereycken et al. (2013) viewed a PSD manifold as a homogeneous space and Massart and Absil (2020) analyzed a quotient geometry on the manifold of PSD matrices. However, it is hard to give a statistical model on these manifolds. More recently, Neuman et al. (2021) proposed a geometry for restricted PSD matrices which has closed-form solutions for many geometric concepts including the Karcher mean. Our paper provides statistical analysis of the Karcher mean corresponding to this geometry.

Distributed PCA To estimate the leading eigenvector, Garber et al. (2017) proposed a sign-fixing averaging approach. To estimate the top K eigenspace, Fan et al. (2019) proposed a projector averaging approach and Charisopoulos et al. (2021) proposed to average local eigenvector matrices after carefully rotating them. Disregarding the information of eigenvalues, both Fan et al. (2019)'s and Charisopoulos et al. (2021)'s methods require the knowledge of the precise location K of a large eigen gap. To alleviate this issue, Bhaskara and Wijewardena (2019) proposed to average the local rank-K approximation matrices and then conduct PCA on the aggregated matrix. Neuman et al. (2021) utilized the same methods as Bhaskara and Wijewardena (2019) except that the average of local rank-K approximation matrices is taken on the manifold of restricted PSD matrices. Neuman et al. (2021) did not provide statistical analysis for their proposed method, while our paper fixes this gap as an application of the main results. Another branch of research turns PCA into the problem of solving a linear system and then solves distributed PCA by some multi-round algorithms. Among them, some make use of the shift-and-invert framework (Garber et al., 2017; Chen et al., 2021), while some use incremental update schemes (Gang et al., 2019; Grammenos et al., 2020; Li et al., 2021).

Notation. By convention, we use regular letters for scalars and bold letters for both vectors and matrices. Given a vector $\boldsymbol{u} \in \mathbb{R}^p$, denote by $\|\boldsymbol{u}\|_2$ its ℓ_2 norm. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times p}$, we use $\|\boldsymbol{A}\|_{\mathrm{F}}$, $\|\boldsymbol{A}\|_2$ and $\|\boldsymbol{A}\|_{\max} = \max_{i,j}|\boldsymbol{A}_{ij}|$ to denote its Frobenius norm, ℓ_2 norm and max norm, respectively. We use $\mathrm{span}(\boldsymbol{A})$ to represent the subspace spanned by the columns of $\boldsymbol{A}$. For a symmetric matrix $\boldsymbol{A}$, denote by $\lambda_j(\boldsymbol{A})$ its jth largest eigenvalue. For two sequences of real numbers $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, we write $a_n \lesssim b_n$ (or $a_n \gtrsim b_n$) if $a_n \leq Cb_n$ (or $a_n \geq Cb_n$) for some constant $C > 0$ independent of n. For an infinitesimal number ϵ, we denote a matrix whose Frobenius norm or max norm is $\mathcal{O}(\epsilon)$ (i.e., $\lesssim \epsilon$) by $\mathcal{O}_{\mathrm{F}}(\epsilon)$ or $\mathcal{O}_{\max}(\epsilon)$, respectively. Given a random variable $x \in \mathbb{R}$, we define $\|x\|_{\psi_2} = \sup_{p\geq 1}(\mathbb{E}|x|^p)^{1/p}/\sqrt{p}$ and $\|x\|_{\psi_1} = \sup_{p\geq 1}(\mathbb{E}|x|^p)^{1/p}/p$. Given two integers $p \geq K > 0$, we denote by $\mathcal{O}_{p\times K}$ the set of matrices in $\mathbb{R}^{p\times K}$ whose columns are orthonormal. Denote by $S(p,K)$ the set of all $p\times p$ PSD matrices of rank K. Denote by $a \vee b = \max\{a,b\}$.

2 The Manifold of Restricted PSD Matrices

In this section, we briefly recap the geometry for restricted PSD matrices (Neuman et al., 2021). To start with, any

PSD matrix $\boldsymbol{A} \in S(p,K)$ has a unique Cholesky decomposition $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top$ such that $\boldsymbol{L} \in \mathbb{R}^{p\times p}$ is a lower triangular matrix and has precisely K positive diagonal elements and $p-K$ zero columns. The jth column of $\boldsymbol{L}$ is zero if and only if the jth column of $\boldsymbol{A}$ is linearly dependent on the previous $j-1$ columns of $\boldsymbol{A}$. Thus, we can rewrite $\boldsymbol{A} = \boldsymbol{N}\boldsymbol{N}^\top$, where $\boldsymbol{N} \in \mathbb{R}^{p\times K}$ consists of K non-zero columns of $\boldsymbol{L}$ without changing the order. Note that $\boldsymbol{N}$ is mock lower triangular, i.e., $\boldsymbol{N}_{ij} = 0$ if $i < j$. We refer to $\boldsymbol{N}$ as the *reduced Cholesky factor* of $\boldsymbol{A}$. To further develop a geometric structure, Neuman et al. (2021) consider the restricted subset $S^*(p,K)$ of $S(p,K)$ such that the first K columns of $\boldsymbol{A} \in S^*(p,K)$ are linearly independent. The set of all reduced Cholesky factors of matrices in $S^*(p,K)$ is denoted by $\mathcal{L}^*(p,K)$, which is equivalent to the set of all mock lower triangular matrices in $\mathbb{R}^{p\times K}$ with positive diagonal elements. Neuman et al. (2021) impose a Riemannian structure on $S^*(p,K)$ and $\mathcal{L}^*(p,K)$ such that the following mappings are isometric,

$$\mathfrak{h} : S^*(p,K) \mapsto \mathcal{L}^*(p,K),\ \boldsymbol{A} \mapsto \boldsymbol{N}, \tag{2.1}$$

$$\mathfrak{g} : \mathcal{L}^*(p,K) \mapsto \mathcal{L}(p,K),\ \boldsymbol{N} \mapsto \boldsymbol{N}', \tag{2.2}$$

where $\boldsymbol{N} = \mathfrak{h}(\boldsymbol{A})$ is the reduced Cholesky factor of $\boldsymbol{A}$, $\mathcal{L}(p,K) = \{\boldsymbol{N}' \in \mathbb{R}^{p\times K} : \boldsymbol{N}'_{ij} = 0, i < j\}$ is endowed with a Euclidean structure, and $\boldsymbol{N}' = \mathfrak{g}(\boldsymbol{N}) \in \mathcal{L}(p,K)$ is defined by $\boldsymbol{N}'_{ii} = \log(\boldsymbol{N}_{ii}), \forall\, i$ and $\boldsymbol{N}'_{ij} = \boldsymbol{N}_{ij}, \forall\, i > j$. We refer to $\boldsymbol{N}' = \mathfrak{g}\circ\mathfrak{h}(\boldsymbol{A})$ as the reduced log-Cholesky factor of $\boldsymbol{A}$. The Karcher mean $\widetilde{\boldsymbol{A}}$ of M restricted PSD matrices $\{\boldsymbol{A}^m\}_{m=1}^M \subset S^*(p,K)$ has a closed-form solution, which is given by

$$\widetilde{\boldsymbol{A}} = \mathfrak{h}^{-1}\circ\mathfrak{g}^{-1}\Big(\frac{1}{M}\sum_{m=1}^M \mathfrak{g}\circ\mathfrak{h}(\boldsymbol{A}^m)\Big). \tag{2.3}$$

The algorithm computing $\widetilde{\boldsymbol{A}}$ is referred to as the Low Rank Cholesky (LRC) algorithm (Neuman et al., 2021).

3 Statistical Analysis of the Karcher Mean

In this section, we provide the first statistical analysis of the Karcher mean under an intrinsic model on the restricted PSD manifold. Then we consider a general signal-plus-noise model under which a deterministic error bound of the Karcher mean is given.

3.1 An intrinsic model

Inspired by the isometry stated in equations (2.1) and (2.2) between the manifold $S^*(p,K)$ of restricted PSD matrices and the Euclidean space $\mathcal{L}(p,K)$, we propose the following intrinsic model. Suppose $\boldsymbol{A} \in S^*(p,K)$ is the signal matrix and denote by $\boldsymbol{N}' = \mathfrak{g}\circ\mathfrak{h}(\boldsymbol{A})$ its reduced log-Cholesky factor. The observations $\{\boldsymbol{A}^m\}_{m=1}^M$ are generated as follows:

$$\boldsymbol{A}^m = \mathfrak{h}^{-1}\circ\mathfrak{g}^{-1}(\boldsymbol{N}' + \boldsymbol{E}^m), \quad m = 1,\dots,M, \tag{3.1}$$

where $\{\boldsymbol{E}^m\}_{m=1}^M \subset \mathcal{L}(p,K)$ are independent and the lower triangular entries of $\boldsymbol{E}^m$ are independent normal variables with mean zero and variance σ^2. Under this intrinsic model, the Karcher mean of $\{\boldsymbol{A}^m\}_{m=1}^M$ can be rewritten as

$$\widetilde{\boldsymbol{A}} = \mathfrak{h}^{-1}\circ\mathfrak{g}^{-1}(\boldsymbol{N}' + \frac{1}{M}\sum_{m=1}^M \boldsymbol{E}^m). \tag{3.2}$$

Using measure concentration, we can obtain a non-asymptotic error bound for the Karcher mean $\widetilde{\boldsymbol{A}}$.

Theorem 3.1 (Intrinsic Model). *Suppose $\boldsymbol{A} \in S^*(p,K)$ is the signal matrix and assume $\|\boldsymbol{A}\|_2 \le C$ for some constant $C > 0$. Assume samples $\{\boldsymbol{A}^m\}_{m=1}^M$ are generated from the intrinsic model (3.1) and denote by $\widetilde{\boldsymbol{A}}$ the Karcher mean of $\{\boldsymbol{A}^m\}_{m=1}^M$. Then there exist some constants $c_1, c_2 > 0$ such that the following inequality*

$$\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_{\mathrm{F}} \le \sqrt{\frac{c_2 pK\sigma^2}{M}} \tag{3.3}$$

holds with probability at least $1 - e^{-c_1 pK}$.

Remark 3.2. *It is worth noting that (3.3) achieves the optimal rate $M^{-1/2}$. In addition, it only depends on the intrinsic dimension $\mathcal{O}(pK)$ of the manifold, which can be much smaller than the ambient dimension p^2.*

3.2 A general signal-plus-noise model

The intrinsic model may be too restricted, so this subsection introduces a general signal-plus-noise model and then provides a deterministic characterization of the Karcher mean. An application of this deterministic error bound to the distributed PCA problem will be given in Section 4. Similar to the intrinsic model, we denote by $\boldsymbol{A} \in S^*(p,K)$ the signal matrix and $\boldsymbol{N} = \mathfrak{h}(\boldsymbol{A})$ its reduced Cholesky factor. The observations $\{\boldsymbol{A}^m\}_{m=1}^M \subset S^*(p,K)$ are given by

$$\boldsymbol{A}^m = (\boldsymbol{N} + \boldsymbol{E}^m)(\boldsymbol{N} + \boldsymbol{E}^m)^\top, \tag{3.4}$$

where $\boldsymbol{E}^m \in \mathbb{R}^{p\times K}$ represents the m-th noise matrix. Here $\boldsymbol{E}^m$ is not necessarily a mock lower triangular matrix, so the model is quite general. Also, the reduced Cholesky factor of $\boldsymbol{A}^m$ is not necessarily $\boldsymbol{N} + \boldsymbol{E}^m$, but rather $(\boldsymbol{N} + \boldsymbol{E}^m)\boldsymbol{Q}^m$ for some orthogonal matrix $\boldsymbol{Q}^m \in \mathcal{O}_{K\times K}$. Denote by $\boldsymbol{N}^m$ the reduced Cholesky factor of $\boldsymbol{A}^m$.

To characterize the Karcher mean (2.3) of $\{\boldsymbol{A}^m\}_{m=1}^M$, we first establish a linear perturbation expansion of QR decomposition below.

Lemma 3.3 (Linear Perturbation Expansion). *Suppose $\boldsymbol{Q} \in \mathcal{O}_{K\times K}$ and $\boldsymbol{R} \in \mathbb{R}^{K\times K}$ is a lower triangular matrix with positive diagonal elements. Given a noise matrix $\boldsymbol{E} \in \mathbb{R}^{K\times K}$, there exist a unique orthogonal matrix $\breve{\boldsymbol{Q}} \in \mathcal{O}_{K\times K}$ and a lower triangular matrix $\breve{\boldsymbol{R}} \in \mathbb{R}^{K\times K}$ with*

non-negative diagonal elements such that $\check{\boldsymbol{R}}\check{\boldsymbol{Q}} = \boldsymbol{R}\boldsymbol{Q}+\boldsymbol{E}$. *When* $\epsilon_0 = \|\boldsymbol{E}\|_{\max}$ *is sufficiently small, we have*

$$\check{\boldsymbol{Q}} = \boldsymbol{Q} + f_{\boldsymbol{R}}(\boldsymbol{E}\boldsymbol{Q}^\top)\boldsymbol{Q} + \mathcal{O}_{\max}(\epsilon_0^2),$$
$$\check{\boldsymbol{R}} = \boldsymbol{R} + \boldsymbol{E}\boldsymbol{Q}^\top - \boldsymbol{R}f_{\boldsymbol{R}}(\boldsymbol{E}\boldsymbol{Q}^\top) + \mathcal{O}_{\max}(\epsilon_0^2),$$

where $f_{\boldsymbol{R}} : \mathbb{R}^{K\times K} \mapsto \mathbb{R}^{K\times K}$ *is given by*

$$f_{\boldsymbol{R}}(\boldsymbol{E}) = \mathcal{U}(\boldsymbol{R}^{-1}\boldsymbol{E}) - (\mathcal{U}(\boldsymbol{R}^{-1}\boldsymbol{E}))^\top,$$
$$\mathcal{U}(\boldsymbol{P})_{ij} = \boldsymbol{P}_{ij}, i < j, \quad \mathcal{U}(\boldsymbol{P})_{ij} = 0, \text{otherwise}.$$

It is worth emphasizing the following properties of $f_{\boldsymbol{R}}$. First, $f_{\boldsymbol{R}}$ is linear in its argument, i.e., $f_{\boldsymbol{R}}(a\boldsymbol{E} + b\boldsymbol{F}) = af_{\boldsymbol{R}}(\boldsymbol{E}) + bf_{\boldsymbol{R}}(\boldsymbol{F})$ for any $a, b \in \mathbb{R}$ and $\boldsymbol{E}, \boldsymbol{F} \in \mathbb{R}^{K\times K}$. Second, $f_{\boldsymbol{R}}(\boldsymbol{E})$ is a skew-symmetric matrix, i.e., $(f_{\boldsymbol{R}}(\boldsymbol{E}))^\top = -f_{\boldsymbol{R}}(\boldsymbol{E})$. Last, $f_{\boldsymbol{R}}$ is bounded in the sense that $\|f_{\boldsymbol{R}}(\cdot)\|_{\mathrm{F}} \le \sqrt{2}\|\boldsymbol{R}^{-1}\|_2\|\cdot\|_{\mathrm{F}}$. Similar first-order perturbation theories exist in the literature for QR, Cholesky, and LU factorization (Chang et al., 1996; Stewart, 1997, 1977; Chang et al., 1997), but none of them provides a linear perturbation expansion with a max-norm control on the remainder term, which is necessary for our development of the error bound on the Karcher mean and the subsequent applications in distributed PCA.

Now we are ready to present a deterministic characterization of the Karcher mean. For convenience, we write $\boldsymbol{N} = (\boldsymbol{R}^\top\ \boldsymbol{B}^\top)^\top$ and $\boldsymbol{E}^m = (\boldsymbol{E}^{1,m^\top}\ \boldsymbol{E}^{2,m^\top})^\top$ such that $\boldsymbol{R}, \boldsymbol{E}^{1,m} \in \mathbb{R}^{K\times K}$ and $\boldsymbol{B}, \boldsymbol{E}^{2,m} \in \mathbb{R}^{(p-K)\times K}$. Note that $\boldsymbol{R}$ is a lower triangular matrix with positive diagonal elements since $\boldsymbol{N} \in \mathcal{L}^*(p, K)$. In the following theorem, we will show that when $\epsilon_0 = \max_m \|\boldsymbol{E}^m\|_{\max}$ is sufficiently small, the reduced Cholesky factor $\widetilde{\boldsymbol{N}}$ of $\widetilde{\boldsymbol{A}}$ differs from $\boldsymbol{N}$ by a term linear in $\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}^m$ and an extra term of order $\mathcal{O}_{\max}(\epsilon_0^2)$. Recall that $S^*(p, K)$ is the manifold of restricted PSD matrices.

Theorem 3.4 (Karcher Mean on $S^*(p, K)$)**.** *When* $\epsilon_0 = \max_m \|\boldsymbol{E}^m\|_{\max}$ *is sufficiently small, the reduced Cholesky factor* $\widetilde{\boldsymbol{N}}$ *of the Karcher mean* $\widetilde{\boldsymbol{A}}$ *of* $\{\boldsymbol{A}^m = (\boldsymbol{N} + \boldsymbol{E}^m)(\boldsymbol{N} + \boldsymbol{E}^m)^\top\}_{m=1}^M$ *on* $S^*(p, K)$ *is*

$$\widetilde{\boldsymbol{N}} = \boldsymbol{N} + \frac{1}{M}\sum_{m=1}^M \boldsymbol{E}^m - \boldsymbol{N}f_{\boldsymbol{R}}(\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}^{1,m}) + \mathcal{O}_{\max}(\epsilon_0^2),$$

where $f_{\boldsymbol{R}}(\cdot)$ *is given in Lemma 3.3.*

From Theorem 3.4, one may easily derive a deterministic upper bound on $\|\widetilde{N} - N\|_{\mathrm{F}}$ using the triangular inequality, which depends on $\|\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}^m\|_{\mathrm{F}}$ and $pK\epsilon_0^2$.

Corollary 3.5. *Under the same conditions of Theorem 3.4, if* $\|\boldsymbol{N}\|_2 \le C$ *and* $\|\boldsymbol{R}^{-1}\| \le C$ *for some constant* $C > 0$*, then we have*

$$\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_F \le \mathcal{O}(\|\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}^m\|_F) + \mathcal{O}(pK\epsilon_0^2).$$

Algorithm 1: LRC-dPCA

Input: $\{\widehat{\boldsymbol{\Sigma}}^m = \frac{1}{n}\sum_i \boldsymbol{x}_i^m \boldsymbol{x}_i^{m\top}\}_{m=1}^M, K$;
Output: $\widetilde{\boldsymbol{V}}$;
Compute $\widehat{\boldsymbol{V}}^m$ and $\widehat{\boldsymbol{\Lambda}}^m$ of $\widehat{\boldsymbol{\Sigma}}^m$ and communicate them to a central server;
Compute the Karcher mean $\widetilde{\boldsymbol{A}}$ of $\widehat{\boldsymbol{V}}^m(\widehat{\boldsymbol{\Lambda}}^m)^2\widehat{\boldsymbol{V}}^{m\top}$ on the manifold of restricted PSD matrices;
Compute the top K eigenspace $\widetilde{\boldsymbol{V}}$ of $\widetilde{\boldsymbol{A}}$.

In applications such as distributed PCA, the Frobenius norm of the average $\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}^m$ is much smaller than that of $\boldsymbol{E}^m$. Thus, by Corollary 3.5, the Karcher mean $\widetilde{\boldsymbol{N}}$ is a better approximation of $\boldsymbol{N}$ than any $\boldsymbol{N}^m$ (the reduced Cholesky factor of $\boldsymbol{A}^m$).

4 Applications to Distributed PCA

This section applies Theorem 3.4 to show that the distributed PCA algorithm, LRC-dPCA proposed by Neuman et al. (2021), achieves the same performance as the full sample PCA when the local sample size is sufficiently large.

4.1 Distributed PCA and LRC-dPCA

We start with the distributed PCA setting as well as the LRC-dPCA algorithm. For simplicity, we consider a balanced setting, in which we have M machines and the m-th machine has n samples $\{\boldsymbol{x}_i^m\}_{i=1}^n \subset \mathbb{R}^p$. Denote by $N = Mn$ the total number of samples. Assume all samples are i.i.d. sub-Gaussian with mean $\boldsymbol{0}$ and covariance $\boldsymbol{\Sigma}$.

Definition 4.1 (sub-Gaussian)**.** *We say a random vector* $\boldsymbol{x} \in \mathbb{R}^p$ *is sub-Gaussian with mean* $\boldsymbol{0}$ *and covariance* $\boldsymbol{\Sigma}$ *if* $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}$ *is sub-Gaussian with mean* $\boldsymbol{0}$ *and covariance* $\boldsymbol{I}_p$*, i.e., there exists a constant* $\sigma > 0$ *such that the following inequality holds,*

$$\mathbb{E}[e^{\lambda\langle \boldsymbol{u}, \boldsymbol{z}\rangle}] \le e^{\frac{\lambda^2\sigma^2}{2}}, \quad \forall\lambda \in \mathbb{R}, \forall \boldsymbol{u} \in \mathbb{R}^p, \|\boldsymbol{u}\|_2 = 1.$$

Remark 4.2. *Fan et al. (2019) and Bhaskara and Wijewardena (2019) use the following* equivalent *definition of a sub-Gaussian vector:* $\boldsymbol{x} \in \mathbb{R}^d$ *is sub-Gaussian with mean* $\boldsymbol{0}$ *and covariance* $\boldsymbol{\Sigma}$ *if there exists a constant* $C > 0$ *such that* $\|\boldsymbol{u}^\top\boldsymbol{x}\|_{\psi_2} \le C\sqrt{\mathbb{E}(\boldsymbol{u}^\top\boldsymbol{x})^2}, \forall \boldsymbol{u} \in \mathbb{R}^d$*. For more information on the equivalent definitions of sub-Gaussian vectors, one may refer to Vershynin (2012).*

Given a positive integer K, the goal is to compute the top K eigenspace of $\boldsymbol{\Sigma}$ using all data on M machines with small communication cost. We consider the LRC-dPCA algorithm proposed by Neuman et al. (2021), collected in Algorithm 1. Following this algorithm, we first compute

$\widehat{\boldsymbol{\Sigma}}^m = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^m \boldsymbol{x}_i^{m\top}$ on each local machine and then compute the top K eigenvectors $\widehat{\boldsymbol{V}}^m = (\widehat{\boldsymbol{v}}_1^m, \ldots, \widehat{\boldsymbol{v}}_K^m) \in \mathcal{O}_{p\times K}$ and eigenvalues $\widehat{\boldsymbol{\Lambda}}^m = \mathrm{diag}(\lambda_1^m, \ldots, \lambda^m)$ of $\widehat{\boldsymbol{\Sigma}}^m$. After communicating these local estimators $\widehat{\boldsymbol{V}}^m, \widehat{\boldsymbol{\Lambda}}^m$ to a central server, we compute the Karcher mean $\widetilde{\boldsymbol{A}}$ of $\{\widehat{\boldsymbol{V}}^m(\boldsymbol{\Lambda}^m)^2\widehat{\boldsymbol{V}}^{m\top}\}_{m=1}^M$ on the manifold of restricted PSD matrices[1]. Finally, the top K eigenvectors $\widetilde{\boldsymbol{V}} \in \mathcal{O}_{p\times K}$ of $\widetilde{\boldsymbol{A}}$ is returned.

4.2 Theoretical analysis

A statistical analysis of the LRC-dPCA algorithm is missing in its original paper (Neuman et al., 2021). In this subsection, we will utilize our deterministic characterization of the Karcher mean on $S^*(p,K)$, i.e., Theorem 3.4, to show that given sufficiently large sub-sample size, LRC-dPCA matches the performance of the full sample PCA. Denote by $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K) \in \mathcal{O}_{p\times K}$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$ the top K eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$, respectively. To ensure the uniqueness of $\mathrm{span}(\boldsymbol{V})$, we assume $\Delta_K = \lambda_K(\boldsymbol{\Sigma}) - \lambda_{K+1}(\boldsymbol{\Sigma}) > 0$. Write $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{A}^\top$ and assume the first K columns of $\boldsymbol{A}$ are linearly independent, i.e., $\boldsymbol{A} \in S^*(p,K)$. When the subsample size n is sufficiently large, we will show that $\widehat{\boldsymbol{A}}^m = \widehat{\boldsymbol{V}}^m(\widehat{\boldsymbol{\Lambda}}^m)^2\widehat{\boldsymbol{V}}^{m\top}$ also belongs to $S^*(p,K)$ with high probability. Here $\widehat{\boldsymbol{V}}^m$ and $\widehat{\boldsymbol{\Lambda}}^m$ denote the top K eigenvectors and eigenvalues of $\widehat{\boldsymbol{\Sigma}}^m$ respectively. Denote by $\widetilde{\boldsymbol{A}}$ the Karcher mean of $\{\widehat{\boldsymbol{A}}^m\}_{m=1}^M$ on $S^*(p,K)$. We further denote by $\boldsymbol{N}, \widehat{\boldsymbol{N}}^m, \widetilde{\boldsymbol{N}}$ the reduced Cholesky factors of $\boldsymbol{A}, \widehat{\boldsymbol{A}}^m, \widetilde{\boldsymbol{A}}$. In addition, we define $\boldsymbol{Q}^* \in \mathcal{O}_{K\times K}$ by the equality $\boldsymbol{N} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{Q}^*$.

In the rest of this subsection, we will apply Theorem 3.4 to study the properties of $\widetilde{\boldsymbol{A}}$. First, we show that $\{\widehat{\boldsymbol{A}}^m\}_{m=1}^M$ follow the general signal-plus-noise model (3.4).

Lemma 4.3. Let $\widehat{\boldsymbol{E}}^m = \widehat{\boldsymbol{\Sigma}}^m\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m\boldsymbol{Q}^* - \boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{Q}^*$, where $\boldsymbol{H}^m = \widehat{\boldsymbol{V}}^{m\top}\boldsymbol{V}$ and $\widehat{\boldsymbol{H}}^m = \mathrm{sgn}(\boldsymbol{H}^m) \overset{\mathrm{def}}{=} \boldsymbol{U}_1\boldsymbol{U}_2^\top$ with $\boldsymbol{U}_1, \boldsymbol{U}_2$ given by the singular value decomposition $\boldsymbol{H}^m = \boldsymbol{U}_1\boldsymbol{\Gamma}\boldsymbol{U}_2^\top$ of $\boldsymbol{H}^m$. Then $\widehat{\boldsymbol{A}}^m = (\boldsymbol{N} + \widehat{\boldsymbol{E}}^m)(\boldsymbol{N} + \widehat{\boldsymbol{E}}^m)^\top$.

Let us make several remarks on $\widehat{\boldsymbol{E}}^m$. It is well-known that

$$\widehat{\boldsymbol{H}}^m = \operatorname*{argmin}_{\boldsymbol{O}\in\mathcal{O}_{K\times K}} \|\widehat{\boldsymbol{V}}^m\boldsymbol{O} - \boldsymbol{V}\|_{\mathrm{F}}$$

and thus $\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m$ is a good estimator of $\boldsymbol{V}$ (Chen et al., 2020). Furthermore, by Lemma H.2, when $\epsilon = \max_m \|\mathcal{E}^m\|_2/\Delta_K \le 1/10$ with $\mathcal{E}^m = \widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}$, $\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m$ has the following first-order expansion around $\boldsymbol{V}$,

$$\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m = \boldsymbol{V} + g(\mathcal{E}^m\boldsymbol{V}) + \mathcal{O}_{\mathrm{F}}(\epsilon^2), \tag{4.1}$$

[1]Here we choose $(\widehat{\boldsymbol{\Lambda}}^m)^2$ rather than $\widehat{\boldsymbol{\Lambda}}^m$ only for technical reasons in the theoretical proofs.

where g is a *linear* function defined in Lemma H.2. Substituting (4.1) into the definition of $\widehat{\boldsymbol{E}}^m$, we obtain the following linear expansion of $\widehat{\boldsymbol{E}}^m$ in terms of $\mathcal{E}^m$,

$$\widehat{\boldsymbol{E}}^m = \mathcal{E}^m\boldsymbol{V}\boldsymbol{Q}^* + \boldsymbol{\Sigma}g(\mathcal{E}^m\boldsymbol{V})\boldsymbol{Q}^* + \mathcal{O}_F(\epsilon^2). \tag{4.2}$$

Since g is linear in its argument, the leading term of $\frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{E}}^m$ is linear in $\frac{1}{M}\sum_{m=1}^M \mathcal{E}^m$. This enables an upper bound for $\|\frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{E}}^m\|_{\mathrm{F}}$, provided by the following lemma.

Lemma 4.4 (Bounding $\|M^{-1}\sum_{m=1}^M \widehat{\boldsymbol{E}}^m\|_{\mathrm{F}}$)**.** *Suppose $\Delta_K > 0$ and $\|\boldsymbol{\Sigma}\|_2$ is bounded. Let $\mathcal{E}^m = \widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}$ and $\epsilon = \max_m \|\mathcal{E}^m\|_2/\Delta_K$. When $\epsilon \le 1/10$, the following bound*

$$\|\frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{E}}^m\|_F \le C\|\frac{1}{M}\sum_{m=1}^M \mathcal{E}^m\|_2 + \mathcal{O}(\epsilon^2)$$

holds for some constant $C > 0$.

To apply Theorem 3.4, we also need to upper bound the max norm $\epsilon_0 = \max_m \|\widehat{\boldsymbol{E}}^m\|_{\max}$. Again, this is based on the first-order expansion (4.2) of $\boldsymbol{E}^m$.

Lemma 4.5 (Bounding $\max_m \|\widehat{\boldsymbol{E}}^m\|_{\max}$)**.** *Assume $\Delta_K > 0$ and $\|\boldsymbol{\Sigma}\|_2$ is bounded. When $\epsilon = \max_m \|\mathcal{E}^m\|_2/\Delta_K \le 1/10$, we have with probability at least $1 - 2Me^{-C_1 n\delta_1^2} - Me^{-C_2\sqrt{\delta_2 n/r}}$ that*

$$\max_m \|\widehat{\boldsymbol{E}}^m\|_{\max} \le C_3\sqrt{\frac{\log(p)}{n}} + \delta_1 + \delta_2,$$

for some constants $C_1, C_2, C_3 > 0$ and $r = \mathrm{Tr}(\boldsymbol{\Sigma})/\lambda_1(\boldsymbol{\Sigma})$. In addition, when $n \gtrsim \log^3(pM)r^2$, we have with probability at least $1 - 2p^{-1}$ that

$$\max_m \|\widehat{\boldsymbol{E}}^m\|_{\max} \le C\sqrt{\frac{\log(pM)}{n}},$$

for some constant $C > 0$.

In Lemma 4.5, we show that $\|\widehat{\boldsymbol{E}}^m\|_{\max} \lesssim \sqrt{\log(p)/n}$ with high probability when $n \gtrsim \log^3(p)r^2$. By (4.2) and Lemma H.1, we can show that $\|\widehat{\boldsymbol{E}}^m\|_{\mathrm{F}} \lesssim \sqrt{p/n}$ with high probability. The upper bound on $\|\widehat{\boldsymbol{E}}^m\|_{\max}$ is thus smaller by a factor of $\sqrt{p/\log(p)}$ than the upper bound on $\|\widehat{\boldsymbol{E}}^m\|_F$. This implies that $\widehat{\boldsymbol{E}}^m$ is delocalized across the entries. Moreover, Lemma 4.5 implies that when we apply Theorem 3.4 to the LRC-dPCA algorithm, the remainder term $\mathcal{O}_{\max}(\epsilon_0^2)$ is negligible compared to the leading term $\frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{E}}^m - \boldsymbol{N}f_{\boldsymbol{R}}(\frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{E}}^{1,m})$. This provides the last key ingredient to the following theorem, which gives an upper bound for $\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_F$. Here $\widetilde{\boldsymbol{N}}$ is the reduced Cholesky factor of the Karcher mean $\widetilde{\boldsymbol{A}}$.

Theorem 4.6 (Bounding $\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_{\mathrm{F}}$)**.** *Assume $\Delta_K > 0$ and $\|\boldsymbol{\Sigma}\|_2$ is bounded. Partition $\boldsymbol{N} = (\boldsymbol{R}^\top\ \boldsymbol{B}^\top)^\top$ such*

that $\boldsymbol{R} \in \mathbb{R}^{K\times K}$ *and* $\boldsymbol{B} \in \mathbb{R}^{(p-K)\times K}$ *and assume* $\|\boldsymbol{R}^{-1}\|_2 \le C$ *for some constant* $C > 0$. *When* $\epsilon = \max_m \|\mathcal{E}^m\|_2/\Delta_K \le 1/10$ *and* $\epsilon_0 = \max_m \|\widehat{\boldsymbol{E}}^m\|_{\max}$ *is sufficiently small, the following bound*

$$\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_F \le \mathcal{O}\left(\|\frac{1}{M}\sum_{m=1}^{M}\mathcal{E}^m\|_2\right) + \mathcal{O}(\epsilon^2) + \mathcal{O}(\sqrt{p}\epsilon_0^2)$$

holds. Define $r = \mathrm{Tr}(\boldsymbol{\Sigma})/\lambda_1(\boldsymbol{\Sigma})$, $\widetilde{r}_1 = (\log^2(pM)r) \vee (\log(pM)\sqrt{p})$ *and* $\widetilde{r}_2 = \sqrt{p}\log^4(pM)r^2$. *Then we have with probability at least* $1 - 4p^{-1}$ *that*

$$\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_F \le \mathcal{O}\left(\frac{\log(p)\sqrt{r}}{\sqrt{Mn}}\right) + \mathcal{O}\left(\frac{\widetilde{r}_1}{n}\right) + \mathcal{O}\left(\frac{\widetilde{r}_2}{n^2}\right).$$

When $n \gtrsim \widetilde{r}_2/\widetilde{r}_1$, *the third term is negligible. When we further assume* $n \gtrsim M\widetilde{r}_1^2/(\log^2(p)r)$, *the upper bound reduces to*

$$\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_F \le \mathcal{O}\left(\frac{\log(p)\sqrt{r}}{\sqrt{Mn}}\right).$$

Theorem 4.6 shows that given sufficiently large local sample size, i.e., $n \gtrsim M\widetilde{r}_1^2/(\log^2(p)r)$, $\widetilde{\boldsymbol{N}}$ is as good as the full sample estimator of $\boldsymbol{N}$ in terms of the Frobenius norm. Moreover, $\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_{\mathrm{F}}$ is of the same order as $\|M^{-1}\sum_{m=1}^{M}\mathcal{E}^m\|_2$ (see Lemma H.1). Note that the singular vectors of $\boldsymbol{N}$ are equal to $\boldsymbol{V}$, the singular values of $\boldsymbol{N}$ are equal to $\boldsymbol{\Lambda}$, and LRC-dPCA uses the singular vectors of $\widetilde{\boldsymbol{N}}$ as an estimator of $\boldsymbol{V}$. Then it follows from Wedin's sin($\boldsymbol{\Theta}$) theorem (Chen et al., 2020) that LRC-dPCA and full sample PCA share the same performance in eigenvector estimation.

Remark 4.7. *Similar to Lemma 4.5, we can show that* $\|\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m - \boldsymbol{V}\|_{\max} \lesssim \sqrt{\log(p)/n}$ *with high probability when* $n \gtrsim \log^3(p)r^2$. *Compared to the upper bound* $\|\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m - \boldsymbol{V}\|_F \lesssim \sqrt{p/n}$, *the max norm bound again implies that the residual matrix* $\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m - \boldsymbol{V}$ *does not concentrate on a few coordinates. This has connections to the infinity norm eigenvector perturbation theory (Fan et al., 2018; Chen et al., 2020; Abbe et al., 2020; Damle and Sun, 2020; Cape et al., 2019). However, most applications in their works require incoherence conditions on the eigenvectors. In contrast, we do not require such conditions.*

4.3 Manifold selection

As one may notice, $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^\top$ may not belong to $S^*(p,K)$, i.e., the first K columns of $\boldsymbol{A}$ may be linearly dependent. If we decompose $\boldsymbol{A} = \boldsymbol{F}\boldsymbol{F}^\top$ for some $\boldsymbol{F} \in \mathbb{R}^{p\times K}$ and write $\boldsymbol{F} = (\boldsymbol{F}_1^\top\ \boldsymbol{F}_2^\top)^\top$ with $\boldsymbol{F}_1 \in \mathbb{R}^{K\times K}$ and $\boldsymbol{F}_2 \in \mathbb{R}^{(p-K)\times K}$. Then the smallest singular value $\sigma_{\min}(\boldsymbol{F}_1)$ of $\boldsymbol{F}_1$ may be zero or very small depending on p. In these cases, the condition $\|\boldsymbol{R}^{-1}\|_2 \le C$ for some constant $C > 0$ in Theorem 4.6 may not hold, and it is not suitable to directly use the manifold $S^*(p,K)$ in the LRC-dPCA algorithm.

Algorithm 2: `find_index` in LRC-dPCA

Input: $\boldsymbol{V}, \boldsymbol{\Lambda}, K$
Output: $\mathcal{I} \subset [p]$
Compute $\boldsymbol{T} = \boldsymbol{V}\boldsymbol{\Lambda}$ and initialize $\mathcal{I} = [0,\dots,0] \in \mathbb{Z}^K$.
for $k = 1$ **to** K **do**
  for $i = 1$ **to** p **do**
    Set $\boldsymbol{T}_k = \boldsymbol{T}[c(\mathcal{I}[1:(k-1)], i), c(1:k)]$.
    Compute score$[i] = \sigma_k(\boldsymbol{T}_k)$.
  end for
  Set $\mathcal{I}[k] = \mathrm{argmax}_i$ score$[i]$.
end for

To fix this issue, we will utilize $\frac{p!}{(p-K)!}$ cousins of the manifold $S^*(p,K)$, or equivalently $\mathcal{L}^*(p,K)$. Let us introduce these cousin manifolds first. Recall that $\mathcal{L}^*(p,K)$ consists of $\boldsymbol{N} \in \mathbb{R}^{p\times K}$ such that $\boldsymbol{N}_{1:K,1:K}$ is a lower triangular matrix with positive diagonal elements. Here $\boldsymbol{N}_{1:K,1:K} \in \mathbb{R}^{K\times K}$ represents the sub-matrix of $\boldsymbol{N}$ with row index $[1,\dots,K]$ and column index $[1,\dots,K]$. Let $\mathcal{I} = [i_1,\dots,i_K]$ be an ordered index set of size K. A cousin $\mathcal{L}_{\mathcal{I}}^*(p,K)$ of $\mathcal{L}^*(p,K)$ consists of $\boldsymbol{N} \in \mathbb{R}^{p\times K}$ such that $\boldsymbol{N}_{\mathcal{I},1:K}$ is a lower triangular matrix with positive diagonal elements. Similarly, we define $S_{\mathcal{I}}^*(p,K)$ as the set of all matrices in $S(p,K)$ with the $\mathcal{I}$-th rows linearly independent. Similar to the relationship between $S^*(p,K)$ and $\mathcal{L}^*(p,K)$, for any $\boldsymbol{A} \in S_{\mathcal{I}}^*(p,K)$, there exists a unique element $\boldsymbol{N} \in \mathcal{L}_{\mathcal{I}}^*(p,K)$ such that $\boldsymbol{A} = \boldsymbol{N}\boldsymbol{N}^\top$. Also, we define the Riemannian structure on $S_{\mathcal{I}}^*(p,K)$ and $\mathcal{L}_{\mathcal{I}}^*(p,K)$ in a way similar to (2.1) and (2.2). In addition, all theory established in Section 3 and 4 can be rephrased in the language of $S_{\mathcal{I}}^*(p,K)$. The only difference is that the row index set $[1,\dots,K]$ is replaced by $\mathcal{I}$.

Now we are in a position to solve the challenge raised at the beginning of this subsection. If $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}^2\boldsymbol{V}^\top$ does not belong to $S^*(p,K)$, then we should choose a suitable ordered index set $\mathcal{I}$ rather than $[1,\dots,K]$, and then apply the LRC-dPCA algorithm on the manifold $S_{\mathcal{I}}^*(p,K)$. Motivated by the condition $\|\boldsymbol{R}^{-1}\| \le C$ in Theorem 4.6, we propose the `find_index` method in Algorithm 2. Given $\boldsymbol{V}$, $\boldsymbol{\Lambda}$, and K, the algorithm outputs an ordered index set $\mathcal{I}$ of size K. To avoid exhaustive search, the algorithm determines $\mathcal{I}$ in a sequential manner. In the kth step, we choose an index $i \in [p]$ such that the k-by-k matrix $\boldsymbol{T}_k = \boldsymbol{T}[c(\mathcal{I}[1:(k-1)], i), c(1:k)]$ has the largest $\sigma_k(\boldsymbol{T}_k)$ among all p candidates, where $c(\cdot)$ indicates the index set. In practice when $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$ is unknown, we can use $\widehat{\boldsymbol{V}}^1$ and $\widehat{\boldsymbol{\Lambda}}^1$ to find a suitable index set and this index set is then shared by all machines.

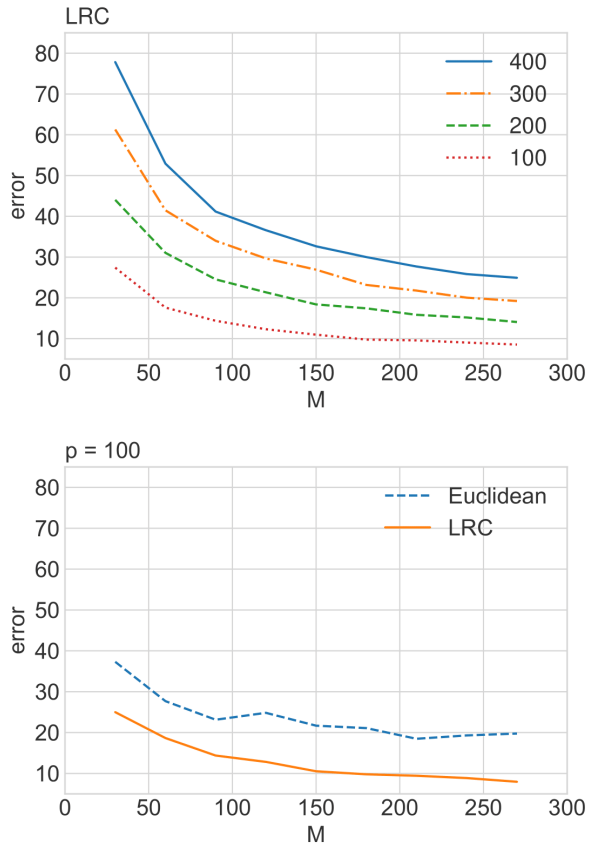

Figure 1: Averaging PSD matrices under the intrinsic model. Top figure: errors $\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_{\mathrm{F}}$ against different M and p with four colored lines labeled by p. Bottom figure: comparisons between LRC and the Euclidean method in terms of $\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_{\mathrm{F}}$ or $\|\widetilde{\boldsymbol{A}}^{\mathrm{eu}} - \boldsymbol{A}\|_{\mathrm{F}}$ against different M.

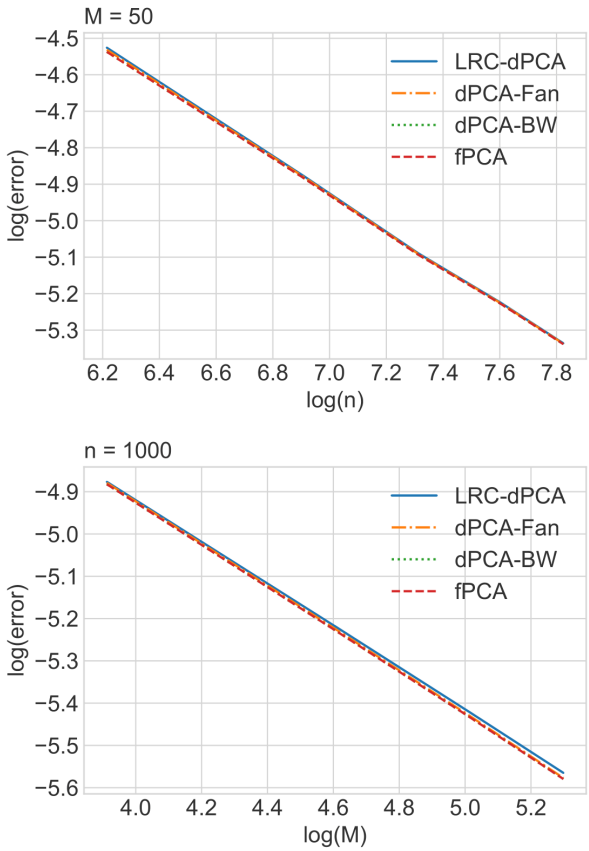

Figure 2: Comparisons of four distributed PCA algorithms, LRC-dPCA, dPCA-Fan, dPCA-BW, fPCA. Top figure: $M = 50$ and $\log(\text{error}) \sim \log(\text{n})$ is reported. Bottom figure: $n = 1000$ and $\log(\text{error}) \sim \log(\text{M})$ is displayed.

5 Numerical Experiments

In this section, we present numerical experiments on three synthetic examples: averaging PSD matrices under the intrinsic model, the distributed PCA problems, and averaging PSD matrices under an extrinsic model.

5.1 Averaging PSD matrices

Our first experiment is to illustrate the concentration of the Karcher mean (2.3) under the intrinsic model (3.1), i.e., Theorem 3.1. We set $K = 5, \sigma^2 = 1$ and let p vary across $[100, 200, 300, 400]$ and let M range from 30 to 270 with an increment of 30. For each p, we generate a $p \times p$ matrix $\boldsymbol{\Sigma}$ with elements i.i.d. $\mathcal{N}(0, 1)$, and then take $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, where $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K)$ and $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_K)$ are the top K left singular vectors and singular values of $\boldsymbol{\Sigma}$, respectively. Given M, we generate $\{\boldsymbol{A}^m\}_{m=1}^M$ from the intrinsic model (3.1). Then the Karcher mean $\widetilde{\boldsymbol{A}}$ of $\{\boldsymbol{A}^m\}_{m=1}^M$ is computed and the error $\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_F$ is reported in the top figure in Figure 1. As our theory shows, the estimation error turns smaller as M increases or p decreases.

In addition, we compare the Karcher mean $\widetilde{\boldsymbol{A}}$, referred to as LRC, with the usual Euclidean method $\widetilde{A}^{\mathrm{eu}}$, which is defined as the best rank-K approximation of $M^{-1}\sum_{m=1}^M \boldsymbol{A}^m$. We take $p = 100$ and repeat the above data generation processing. The errors $\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_{\mathrm{F}}$ and $\|\widetilde{\boldsymbol{A}}^{\mathrm{eu}} - \boldsymbol{A}\|_{\mathrm{F}}$ are reported in the bottom figure in Figure 1. As displayed in the figure, under the intrinsic model, the geometry-aware method, LRC, outperforms the Euclidean method. This justifies the intuition that for models with specific geometric structures, it is better to take that geometric information into account.

5.2 Distributed PCA

Our second experiment studies the Karcher mean under a general signal-plus-noise model. Specifically, we consider the distributed PCA problems and numerically verify Theorem 4.6, which shows that LRC-dPCA achieves the same

performance as full sample PCA (fPCA). In our setting, $p = 100$, $K = 5$, the population covariance $\boldsymbol{\Sigma}$ is generated by $\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{V}^\top + 0.3\boldsymbol{I}_p$, where $\boldsymbol{V} \in \mathbb{R}^{p\times K}$ with elements i.i.d. $\mathcal{N}(0,1)$. We first fix the number of machines $M = 50$ and let the sub-sample size n vary across $[500, 1000, \ldots, 2500]$. On the m-th machine, we generate n i.i.d. samples $\{\boldsymbol{x}_i^m\}_{i=1}^n$ from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and compute the local sample covariance matrix $\widehat{\boldsymbol{\Sigma}}^m = \sum_{i=1}^n \boldsymbol{x}_i^m \boldsymbol{x}_i^{m\top}/n$. Then we apply four methods , namely fPCA, LRC-dPCA, dPCA-Fan (Fan et al., 2019), and dPCA-BW (Bhaskara and Wijewardena, 2019), to compute the top K eigenvectors of $\boldsymbol{\Sigma}$. Let $\widehat{\boldsymbol{V}} \in \mathcal{O}_{p\times K}$ be the estimated top K eigenvectors. The error is defined as $\|\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top - \boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top\|_F$, which is the distance between the population projection matrix $\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{V})^{-1}\boldsymbol{V}^\top$ and the estimated projection matrix $\widehat{\boldsymbol{V}}\widehat{\boldsymbol{V}}^\top$. For each n and each method, the experiment is repeated 100 times and the average of error is recorded. The top figure in Figure 2 displays the relationship between $\log(\text{error})$ and $\log(n)$ for all methods. It turns out that all methods share similar performance and there is a linear relationship between $\log(\text{error})$ and $\log(n)$ with slope $-1/2$, which verifies the relationship $\text{error} \sim \text{n}^{-1/2}$. Next, we fix the sub-sample size $n = 1000$ and let M vary across $[50, 100, \ldots, 200]$ and repeat the above procedures. The relationship between $\log(\text{error})$ and $\log(M)$ is reported in the bottom figure of Figure 2. As it displayed, all four methods are almost the same and there is also a linear relationship between $\log(\text{error})$ and $\log(M)$ with slope $-1/2$, which indicates $\text{error} \sim \text{M}^{-1/2}$. Since there is no specific geometric information in the setting, it is expected that LRC-dPCA only matches (rather than surpasses) the performance of the state-of-the-art methods, dPCA-Fan, dPCA-BW, and the optimal method, fPCA.

5.3 Averaging PSD matrices (extrinsic)

Our third experiment considers another signal-plus-noise model, which adds extrinsic noises to the intrinsic model. Specifically, we set $p = 100$, $K = 5$, and we generate a $p \times p$ matrix $\boldsymbol{\Sigma}$ with elements i.i.d. $\mathcal{N}(0,1)$, and then take $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top$, where $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K)$ and $\boldsymbol{\Lambda} = (\lambda_1, \ldots, \lambda_K)$ are the top K left singular vectors and singular values of $\boldsymbol{\Sigma}$, respectively. Given M and σ^2, we generate $\{\boldsymbol{A}^m\}_{m=1}^M$ from the intrinsic model (3.1). Then we add extrinsic noises to $\boldsymbol{A}^m$ as follows. For each m, we generate $\{\boldsymbol{x}_i^m\}_{i=1}^{2000}$ i.i.d. from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{A}^m + 0.01\boldsymbol{I}_p)$, compute $\widehat{\boldsymbol{\Sigma}}^m = \sum_{i=1}^{2000} \boldsymbol{x}_i^m \boldsymbol{x}_i^{m\top}/2000$, and set $\boldsymbol{A}'^m$ as the best rank-K approximation of $\widehat{\boldsymbol{\Sigma}}^m$. We compute the Karcher mean $\widetilde{\boldsymbol{A}}$ of $\{\boldsymbol{A}'^m\}_{m=1}^M$, which is referred to as LRC, and report the error $\|\widetilde{\boldsymbol{A}} - \boldsymbol{A}\|_{\text{F}}$. In contrast, we also apply the Euclidean method, which computes the best rank-K approximation $\widetilde{\boldsymbol{A}}^{\text{eu}}$ of $\sum_{m=1}^M \boldsymbol{A}'^m/M$, and report the error $\|\widetilde{\boldsymbol{A}}^{\text{eu}} - \boldsymbol{A}\|_{\text{F}}$. First, we set $\sigma^2 = 0.5$ and let M vary across $[100, 200, \ldots, 1000]$. The errors of both methods are displayed in the top figure of Figure 3. As shown in the figure, the geometry-aware method, LRC, still outperforms the Euclidean method even if extrinsic noises are added to the intrinsic model. Next, we fix $M = 400$ and let σ^2 range from $[0, 0.1 \ldots, 0.7]$. The errors of both methods are shown in the bottom figure of Figure 3. Recall that σ^2 denotes the strength of intrinsic noises. The bottom figure indicates that when the intrinsic noises are small, then the geometry-aware method and the Euclidean method are comparable, but when the intrinsic noises becomes large, the geometry-aware method tends to outperform the Euclidean method. Overall, this experiment shows that, in a general signal-plus-noise model, if there exist large intrinsic noises, then it is better to utilize the geometry-aware method.

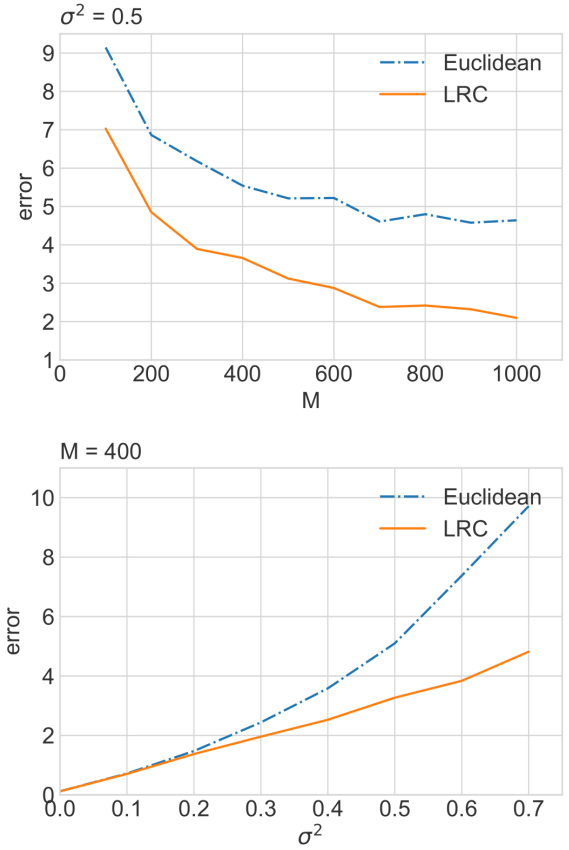

Figure 3: Comparisons of LRC and the Euclidean method in averaging PSD matrices under an extrinsic model. Top figure: $\sigma^2 = 0.5$ and the error against M is reported. Bottom figure: $M = 400$ and the error versus σ^2 is displayed.

6 Concluding Remarks

This paper considers the geometry of restricted PSD matrices proposed by Neuman et al. (2021). In particular, we provide a non-asymptotic statistical analysis of the Karcher mean of restricted PSD matrices under an intrinsic model.

Moreover, for general signal-plus-noise models, we establish a deterministic error bound concerning the Karcher mean. This is based on a linear perturbation expansion of the QR decomposition, which may be of independent interest. As an application, we use the deterministic error analysis of the Karcher mean to prove that the distributed PCA algorithm, LRC-dPCA, achieves the same performance as the full sample PCA. Motivated by the established theory, we propose a manifold selection procedure for the LRC-dPCA algorithm. Finally, we carry out three synthetic numerical experiments to verify our theories. One observation in the experiment is that if data model has certain geometric structure, then it is better to utilize the geometry-aware method.

Several interesting topics are worth of future studiesying. In manifold-valued data analysis (Patrangenaru and Ellingson, 2016), it is remains to determine which statistical model is more suitable for the given data. For example, the highly anisotropic diffusion tensor images are modelled as PSD matrices (Bonnabel et al., 2013), so it is interesting to investigate the performances of the proposed intrinsic model for such data. Second, it is interesting to extend our study to regression, classification, and clustering problems.

References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of statistics*, 48(3):1452.

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM journal on matrix analysis and applications*, 29(1):328–347.

Bhaskara, A. and Wijewardena, P. M. (2019). On distributed averaging for stochastic k-pca. In *Advances in Neural Information Processing Systems*, volume 32.

Bhattacharya, R. and Patrangenaru, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. *The Annals of Statistics*, 31(1):1–29.

Bhattacharya, R. and Patrangenaru, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds—ii. *The Annals of Statistics*, 33(3):1225–1259.

Bigot, J. and Gendre, X. (2013). Minimax properties of fréchet means of discretely sampled curves. *The Annals of Statistics*, 41(2):923–956.

Bini, D. A. and Iannazzo, B. (2013). Computing the karcher mean of symmetric positive definite matrices. *Linear Algebra and its Applications*, 438(4):1700–1710.

Bonnabel, S., Collard, A., and Sepulchre, R. (2013). Rank-preserving geometric means of positive semi-definite matrices. *Linear Algebra and its Applications*, 438(8):3202–3216.

Bonnabel, S. and Sepulchre, R. (2010). Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1055–1070.

Cape, J., Tang, M., and Priebe, C. E. (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *The Annals of Statistics*, 47(5):2405–2439.

Chang, X.-W., Paige, C. C., and Stewart, G. (1996). New perturbation analyses for the cholesky factorization. *IMA journal of numerical analysis*, 16(4):457–484.

Chang, X.-W., Paige, C. C., and Stewart, G. (1997). Perturbation analyses for the qr factorization. *SIAM Journal on Matrix Analysis and Applications*, 18(3):775–791.

Charisopoulos, V., Benson, A. R., and Damle, A. (2021). Communication-efficient distributed eigenspace estimation. *SIAM Journal on Mathematics of Data Science*, 3(4):1067–1092.

Chen, X., Lee, J. D., Li, H., and Yang, Y. (2021). Distributed estimation for principal component analysis: An enlarged eigenspace analysis. *Journal of the American Statistical Association*, pages 1–12.

Chen, Y., Chi, Y., Fan, J., and Ma, C. (2020). Spectral methods for data science: A statistical perspective. *arXiv preprint arXiv:2012.08496*.

Cornea, E., Zhu, H., Kim, P., Ibrahim, J. G., and Initiative, A. D. N. (2017). Regression models on riemannian symmetric spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):463–482.

Damle, A. and Sun, Y. (2020). Uniform bounds for invariant subspace perturbations. *SIAM Journal on Matrix Analysis and Applications*, 41(3):1208–1236.

Dryden, I. L., Koloydenko, A., and Zhou, D. (2009). Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123.

Fan, J., Wang, D., Wang, K., and Zhu, Z. (2019). Distributed estimation of principal eigenspaces. *Annals of statistics*, 47(6):3009.

Fan, J., Wang, W., and Zhong, Y. (2018). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42.

Faraki, M., Harandi, M. T., and Porikli, F. (2016). Image set classification by symmetric positive semi-definite matrices. In *2016 IEEE Winter conference on applications of computer vision (WACV)*, pages 1–8. IEEE.

Gang, A., Raja, H., and Bajwa, W. U. (2019). Fast and communication-efficient distributed pca. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7450–7454. IEEE.

Garber, D., Shamir, O., and Srebro, N. (2017). Communication-efficient algorithms for distributed stochastic principal component analysis.

Grammenos, A., Mendoza Smith, R., Crowcroft, J., and Mascolo, C. (2020). Federated principal component analysis. *Advances in Neural Information Processing Systems*, 33.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Journée, M., Bach, F., Absil, P.-A., and Sepulchre, R. (2010). Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351.

Karcher, H. (1977). Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541.

Li, X., Wang, S., Chen, K., and Zhang, Z. (2021). Communication-efficient distributed svd via local power iterations. In *International Conference on Machine Learning*, pages 6504–6514. PMLR.

Mackey, L., Talwalkar, A., and Jordan, M. I. (2011). Divide-and-conquer matrix factorization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 1134–1142.

Massart, E. and Absil, P.-A. (2020). Quotient geometry with simple geodesics for the manifold of fixed-rank positive-semidefinite matrices. *SIAM Journal on Matrix Analysis and Applications*, 41(1):171–198.

Moakher, M. (2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 26(3):735–747.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Neuman, A. M., Xie, Y., and Sun, Q. (2021). Restricted riemannian geometry for positive semidefinite matrices. *arXiv preprint arXiv:2105.14691*.

Patrangenaru, V. and Ellingson, L. (2016). *Nonparametric statistics on manifolds and their applications to object data analysis*. CRC Press, Taylor & Francis Group Boca Raton.

Pilanci, M. and Wainwright, M. J. (2015). Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115.

Stewart, G. (1977). Perturbation bounds for the qr factorization of a matrix. *SIAM Journal on Numerical Analysis*, 14(3):509–518.

Stewart, G. (1997). On the perturbation of lu and cholesky factors. *IMA Journal of Numerical Analysis*, 17(1):1–6.

Vandereycken, B., Absil, P.-A., and Vandewalle, S. (2013). A riemannian geometry with complete geodesics for the set of positive semidefinite matrices of fixed rank. *IMA Journal of Numerical Analysis*, 33(2):481–514.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, R., Guo, H., Davis, L. S., and Dai, Q. (2012). Covariance discriminative learning: A natural and efficient approach to image set classification. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2496–2503. IEEE.

APPENDIX

A Proof of Theorem 3.1

Proof of Theorem 3.1. Recall that the Karcher mean $\widetilde{\boldsymbol{A}}$ of $\{\boldsymbol{A}^m\}_{m=1}^M$ under the intrinsic model is given by (3.1). First, we give an upper bound on the Frobenius norm of $\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}^m$. By the intrinsic model, we know $\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}^m$ is a mock lower triangular matrix with lower triangular elements i.i.d. $\mathcal{N}(0,\sigma^2/M)$. Therefore, by the concentration of χ^2 ((2.19) in Wainwright (2019)), for all $t\in(0,1)$, we have

$$\|\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}^m\|_{\mathrm{F}}^2 \le \frac{pK\sigma^2}{M}(1+t), \tag{A.1}$$

with probability at least $1-e^{-pKt^2/8}$. In a similar spirit, using union bound, we have for $t\in(0,1)$,

$$\max_{i=1,\ldots,K}|\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}_{ii}^m|^2 \le \frac{\sigma^2}{M}(1+t) \tag{A.2}$$

with probability at least $1-Ke^{-t^2/8}$. Here $\boldsymbol{E}_{ii}^m$ is the (i,i)-th element of $\boldsymbol{E}^m$. Thus with high probability, $\max_{i=1,\ldots,K}|\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}_{ii}^m|^2 \le 1/2$ and

$$|\exp(\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}_{ii}^m)-1| \le 2|\frac{1}{M}\sum_{m=1}^M \boldsymbol{E}_{ii}^m|, \tag{A.3}$$

where we use the inequality $|\exp(x)-1|\le 2|x|$ for $x\le 1/2$. Denote by $\boldsymbol{N}$ the reduced Cholesky factor of $\boldsymbol{A}$. Then it holds that $\|\boldsymbol{N}\|_2 = \|\boldsymbol{A}\|_2^{1/2} \le C^{1/2}$ for some constant $C>0$. Furthermore, by (3.1), we have for some constants $c_1,c_2>0$ that

$$\|\widetilde{\boldsymbol{A}}-\boldsymbol{A}\|_F \le \sqrt{\frac{c_2pK\sigma^2}{M}}, \tag{A.4}$$

with probability at least $1-e^{-c_1pK}$. □

B Proof of Lemma 3.3

Proof of Lemma 3.3. The proof of this lemma is split up into three steps. First, we assume $\boldsymbol{Q}=\boldsymbol{I}_K$ and show that $\breve{\boldsymbol{Q}}$ has the form of $\boldsymbol{I}_K+\breve{\boldsymbol{P}}+\mathcal{O}_{\max}(\epsilon_0^2)$, where $\breve{\boldsymbol{P}}\in\mathbb{R}^{K\times K}$ is a skew-symmetric matrix of order $\mathcal{O}_{\max}(\epsilon_0)$. Second, by taking upper triangular off-diagonal elements of $(\boldsymbol{R}+\boldsymbol{E})\breve{\boldsymbol{Q}}^\top$ as zero, we derive a closed-form expression of $\breve{\boldsymbol{P}}$ (up to a higher-order term). Motivated by this closed-form expression, we define a function $f_{\boldsymbol{R}}:\mathbb{R}^{K\times K}\mapsto\mathbb{R}^{K\times K}$ satisfying several desired conditions. For example, we have $\breve{\boldsymbol{Q}}=\boldsymbol{I}_K+f_{\boldsymbol{R}}(\boldsymbol{E})+\mathcal{O}_{\max}(\epsilon_0^2)$ and $f_{\boldsymbol{R}}$ is linear in its argument. Third, we extend the results to the general case when $\boldsymbol{Q}\in\mathcal{O}_{K\times K}$ may differ from $\boldsymbol{I}_K$.

Step 1. When $\epsilon_0=\|\boldsymbol{E}\|_{\max}$ is sufficiently small, the matrix $\boldsymbol{R}+\boldsymbol{E}$ is still non-singular and by QR decomposition there exists a unique orthogonal matrix $\breve{\boldsymbol{Q}}\in\mathcal{O}_{K\times K}$ such that $\breve{\boldsymbol{R}}=(\boldsymbol{R}+\boldsymbol{E})\breve{\boldsymbol{Q}}^\top$ is a lower triangular matrix with positive diagonal elements. In this step, we will show that $\breve{\boldsymbol{Q}}$ has a form of $\boldsymbol{I}_K+\breve{\boldsymbol{P}}+\mathcal{O}_{\max}(\epsilon_0^2)$, where $\breve{\boldsymbol{P}}\in\mathbb{R}^{K\times K}$ is a skew-symmetric matrix of order $\mathcal{O}_{\max}(\epsilon_0)$. To that end, we construct $\breve{\boldsymbol{Q}}$ as a product of $K(K-1)/2$ rotation matrices $\{\breve{\boldsymbol{Q}}^{ij},1\le i<j\le K\}$, which set the upper triangular off-diagonal elements as zero in a sequential fashion. In specific, we arrange these $K(K-1)/2$ rotation matrices in a prescribed order, i.e., $\{(1,2),\ldots,(1,K),(2,3),\ldots,(K-1,K)\}$. In this way, we may relabel $\{\breve{\boldsymbol{Q}}^{ij},1\le i<j\le K\}$ as $\{\breve{\boldsymbol{Q}}^{(s)},1\le s\le K(K-1)/2\}$ and write $\breve{\boldsymbol{Q}}=\breve{\boldsymbol{Q}}^{(K(K-1)/2)}\cdots\breve{\boldsymbol{Q}}^{(1)}$. In the remainder of this proof, we will use $s(i,j)$ to represent the s-index of the (i,j)th rotation matrix $\breve{\boldsymbol{Q}}^{ij}$.

For each (i,j), we set the rotation matrix $\breve{\boldsymbol{Q}}^{ij}$ as

$$\breve{\boldsymbol{Q}}_{ii}^{ij}=\breve{\boldsymbol{Q}}_{jj}^{ij}=\cos(\theta^{ij}),\quad \breve{\boldsymbol{Q}}_{ij}^{ij}=-\breve{\boldsymbol{Q}}_{ji}^{ij}=\sin(\theta^{ij}),\quad \breve{\boldsymbol{Q}}_{kk}^{ij}=1,\ \forall k\ne i,j,\quad \breve{\boldsymbol{Q}}_{kl}^{ij}=0,\ \text{otherwise},$$

where θ^{ij} is chosen in a sequential fashion such that the (i,j)th element of $\breve{\boldsymbol{R}}^{(s(i,j))} := (\boldsymbol{R}+\boldsymbol{E})\breve{\boldsymbol{Q}}^{(1)^\top}\cdots\breve{\boldsymbol{Q}}^{(s(i,j))^\top}$ is zero and the diagonal elements of $\breve{\boldsymbol{R}}^{(s(i,j))}$ keep positive. Note that $\breve{\boldsymbol{Q}}^{ij}$ is by definition an orthogonal matrix. A simple calculation gives that $\theta^{ij} = \theta^{(s(i,j))} = \arctan(\breve{\boldsymbol{R}}_{ij}^{(s(i,j)-1)}/\breve{\boldsymbol{R}}_{ii}^{(s(i,j)-1)})$.

Next, we show that θ^{ij} is a small quantity of order $\mathcal{O}(\epsilon_0)$ via an deductive argument. First, when $\epsilon_0 = \|\boldsymbol{E}\|_{\max}$ is sufficiently small, $\theta^{(1)} = \theta^{12} = \arctan(\boldsymbol{E}_{12}/(\boldsymbol{R}_{11}+\boldsymbol{E}_{11}))$ is a small quantity of order $\mathcal{O}(\epsilon_0)$. Thus, by definition of $\breve{\boldsymbol{Q}}^{12}$, we have $\breve{\boldsymbol{Q}}^{(1)} = \breve{\boldsymbol{Q}}^{12} = \boldsymbol{I}_K + \mathcal{O}_{\max}(\epsilon_0)$ and

$$\breve{\boldsymbol{R}}^{(1)} = (\boldsymbol{R}+\boldsymbol{E})(\boldsymbol{I}_K + \breve{\boldsymbol{Q}}^{(1)^\top} - \boldsymbol{I}_K) = \boldsymbol{R}+\boldsymbol{E}+\boldsymbol{R}(\breve{\boldsymbol{Q}}^{(1)^\top}-\boldsymbol{I}_K)+\mathcal{O}_{\max}(\epsilon_0^2).$$

Note that the error matrix $\boldsymbol{E}^{(1)} := \breve{\boldsymbol{R}}^{(1)} - \boldsymbol{R}$ is again of order $\mathcal{O}_{\max}(\epsilon_0)$. This implies that $\theta^{(2)} = \theta^{13} = \arctan(\boldsymbol{E}_{13}^{(1)}/(\boldsymbol{R}_{11}+\boldsymbol{E}_{11}^{(1)}))$ is also a small quantity of order $\mathcal{O}(\epsilon_0)$. Applying this deductive argument $K(K-1)/2$ times, we conclude that all $\theta^{(s)}$, $1 \le s \le K(K-1)/2$, are small quantities of order $\mathcal{O}(\epsilon_0)$.

Now we are able to show that $\breve{\boldsymbol{Q}}$ has a form of $\boldsymbol{I}_K + \breve{\boldsymbol{P}} + \mathcal{O}_{\max}(\epsilon_0^2)$, where $\breve{\boldsymbol{P}} \in \mathbb{R}^{K\times K}$ is a skew-symmetric matrix of order $\mathcal{O}_{\max}(\epsilon_0)$. Since θ^{ij} is of order $\mathcal{O}(\epsilon_0)$, by Taylor expansion, we have $\sin(\theta^{ij}) = \theta^{ij} + \mathcal{O}(\epsilon_0^3)$ and $1-\cos(\theta^{ij}) = \mathcal{O}(\epsilon_0^2)$. Thus, we can rewrite $\breve{\boldsymbol{Q}}^{ij}$ as

$$\breve{\boldsymbol{Q}}^{ij} = \boldsymbol{I}_K + \breve{\boldsymbol{P}}^{ij} + \mathcal{O}_{\max}(\epsilon_0^2),$$

where $\breve{\boldsymbol{P}}_{ij}^{ij} = -\breve{\boldsymbol{P}}_{ij}^{ij} = \theta^{ij}$ and $\breve{\boldsymbol{P}}_{kl}^{ij} = 0$ otherwise. Since $\breve{\boldsymbol{P}}^{ij}$ is of order $\mathcal{O}_{\max}(\epsilon_0)$, we have

$$\begin{aligned}
\breve{\boldsymbol{Q}} &= \breve{\boldsymbol{Q}}^{(K(K-1)/2)}\cdots\breve{\boldsymbol{Q}}^{(1)} \\
&= (\boldsymbol{I}_K + \breve{\boldsymbol{P}}^{(K(K-1)/2)} + \mathcal{O}_{\max}(\epsilon_0^2))\cdots(\boldsymbol{I}_K + \breve{\boldsymbol{P}}^{(1)} + \mathcal{O}_{\max}(\epsilon_0^2)) \\
&= \boldsymbol{I}_K + \sum_{s=1}^{K(K-1)/2} \breve{\boldsymbol{P}}^{(s)} + \mathcal{O}_{\max}(\epsilon_0^2) \\
&= \boldsymbol{I}_K + \breve{\boldsymbol{P}} + \mathcal{O}_{\max}(\epsilon_0^2), \qquad \text{(B.1)}
\end{aligned}$$

where $\breve{\boldsymbol{P}}^{(s(i,j))} = \breve{\boldsymbol{P}}^{ij}$ and $\breve{\boldsymbol{P}} = \sum_{s=1}^{K(K-1)/2}\breve{\boldsymbol{P}}^{(s)}$. Since $\breve{\boldsymbol{P}}^{(s)}$ is skew-symmetric and of order $\mathcal{O}_{\max}(\epsilon_0)$ for all s, $\breve{\boldsymbol{P}}$ is also a skew-symmetric matrix of order $\mathcal{O}_{\max}(\epsilon_0)$, which concludes the proof of step 1.

Step 2. Now we are ready to derive a closed-form expression of $\breve{\boldsymbol{P}}$ (maybe up to a higher-order term) by taking upper triangular off-diagonal elements of $(\boldsymbol{R}+\boldsymbol{E})\breve{\boldsymbol{Q}}^\top$ as zero. Substituting (B.1) into $(\boldsymbol{R}+\boldsymbol{E})\breve{\boldsymbol{Q}}^\top$, we obtain

$$\begin{aligned}
(\boldsymbol{R}+\boldsymbol{E})\breve{\boldsymbol{Q}}^\top &= \boldsymbol{R}+\boldsymbol{E}+\boldsymbol{R}\breve{\boldsymbol{P}}^\top + \mathcal{O}_{\max}(\epsilon_0^2) \\
&= \boldsymbol{R}+\boldsymbol{E}-\boldsymbol{R}\breve{\boldsymbol{P}} + \mathcal{O}_{\max}(\epsilon_0^2), \qquad \text{(B.2)}
\end{aligned}$$

where the second equality follows from the skew-symmetry of $\breve{\boldsymbol{P}}$. Since $\boldsymbol{R}$ is a lower triangular matrix with positive diagonal elements, $\boldsymbol{R}$ is invertible and $\boldsymbol{R}^{-1}$ is also a lower triangular matrix. As a result, the matrix $\boldsymbol{R}^{-1}(\boldsymbol{R}+\boldsymbol{E})\breve{\boldsymbol{Q}}^\top$ is also a lower triangular matrix. Multiplying LHS and RHS of (B.2) by $\boldsymbol{R}^{-1}$ simultaneously, we obtain

$$\boldsymbol{R}^{-1}(\boldsymbol{R}+\boldsymbol{E})\breve{\boldsymbol{Q}}^\top = \boldsymbol{I}_K + \boldsymbol{R}^{-1}\boldsymbol{E} - \breve{\boldsymbol{P}} + \mathcal{O}_{\max}(\epsilon_0^2). \qquad \text{(B.3)}$$

For convenience, we define a function $\mathcal{U}(\cdot) : \mathbb{R}^{K\times K} \mapsto \mathbb{R}^{K\times K}, \boldsymbol{P} \mapsto \mathcal{U}(\boldsymbol{P})$, where $\mathcal{U}(\boldsymbol{P})$ takes the upper triangular off-diagonal elements of $\boldsymbol{P}$, i.e.,

$$\mathcal{U}(\boldsymbol{P})_{ij} = \boldsymbol{P}_{ij},\ i<j, \quad \mathcal{U}(\boldsymbol{P})_{ij} = 0, \text{ otherwise.}$$

Since $\boldsymbol{R}^{-1}(\boldsymbol{R}+\boldsymbol{E})\breve{\boldsymbol{Q}}^\top$ is a lower triangular matrix, we have by (B.3) that

$$\mathcal{U}(\breve{\boldsymbol{P}}) = \mathcal{U}(\boldsymbol{R}^{-1}\boldsymbol{E}) + \mathcal{O}_{\max}(\epsilon_0^2).$$

Since $\breve{\boldsymbol{P}}$ is skew-symmetric, we get the following closed-form solution of $\breve{\boldsymbol{P}}$ (up to a higher-order term),

$$\breve{\boldsymbol{P}} = \mathcal{U}(\boldsymbol{R}^{-1}\boldsymbol{E}) - (\mathcal{U}(\boldsymbol{R}^{-1}\boldsymbol{E}))^\top + \mathcal{O}_{\max}(\epsilon_0^2).$$

Motivated by the linear expansion of $\check{\boldsymbol{P}}$, we define the following function,

$$f_{\boldsymbol{R}}: \mathbb{R}^{K\times K} \mapsto \mathbb{R}^{K\times K}, \quad \boldsymbol{E} \mapsto f_{\boldsymbol{R}}(\boldsymbol{E}) \coloneqq \mathcal{U}(\boldsymbol{R}^{-1}\boldsymbol{E}) - (\mathcal{U}(\boldsymbol{R}^{-1}\boldsymbol{E}))^\top.$$

Note that $f_{\boldsymbol{R}}$ is linear in the sense that $f_{\boldsymbol{R}}(a\boldsymbol{E}+b\boldsymbol{F}) = af_{\boldsymbol{R}}(\boldsymbol{E}) + bf_{\boldsymbol{R}}(\boldsymbol{F})$ for all $a, b \in \mathbb{R}$ and $\boldsymbol{E}, \boldsymbol{F} \in \mathbb{R}^{K\times K}$. Also, the image $f_{\boldsymbol{R}}(\boldsymbol{E})$ is a skew-symmetric matrix, i.e., $(f_{\boldsymbol{R}}(\boldsymbol{E}))^\top = -f_{\boldsymbol{R}}(\boldsymbol{E})$. Moreover, we have $\|f_{\boldsymbol{R}}(\boldsymbol{E})\|_{\mathrm{F}} \le \sqrt{2}\|\boldsymbol{R}^{-1}\|_2\|\boldsymbol{E}\|_{\mathrm{F}}$. By (B.1), we can rewrite $\check{\boldsymbol{Q}}$ as follows,

$$\check{\boldsymbol{Q}} = \boldsymbol{I}_K + f_{\boldsymbol{R}}(\boldsymbol{E}) + \mathcal{O}_{\max}(\epsilon_0^2). \tag{B.4}$$

Moreover, by definition of $\check{\boldsymbol{R}}$, we have

$$\check{\boldsymbol{R}} = (\boldsymbol{R}+\boldsymbol{E})\check{\boldsymbol{Q}}^\top = \boldsymbol{R} + \boldsymbol{E} - \boldsymbol{R}f_{\boldsymbol{R}}(\boldsymbol{E}) + \mathcal{O}_{\max}(\epsilon_0^2).$$

Step 3. In general, when $\boldsymbol{Q} \in \mathcal{O}_{K\times K}$ may differ from $\boldsymbol{I}_K$, we can transform the QR decomposition $\check{\boldsymbol{R}}\check{\boldsymbol{Q}} = \boldsymbol{R}\boldsymbol{Q} + \boldsymbol{E}$ suitably and apply the results in the previous two steps to prove the lemma. In specific, we have

$$\check{\boldsymbol{R}}\check{\boldsymbol{Q}}\boldsymbol{Q}^\top = \boldsymbol{R} + \boldsymbol{E}\boldsymbol{Q}^\top.$$

When $\epsilon_0 = \|\boldsymbol{E}\|_{\max}$ is sufficiently small, $\|\boldsymbol{E}\boldsymbol{Q}^\top\|_{\max} \le \sqrt{K}\epsilon_0$ can also be sufficiently small. In addition, $\check{\boldsymbol{Q}}\boldsymbol{Q}^\top$ is still an orthogonal matrix that appears in the QR decomposition of $\boldsymbol{R} + \boldsymbol{E}\boldsymbol{Q}^\top$. Therefore, by (B.4), we have

$$\check{\boldsymbol{Q}}\boldsymbol{Q}^\top = \boldsymbol{I}_K + f_{\boldsymbol{R}}(\boldsymbol{E}\boldsymbol{Q}^\top) + \mathcal{O}_{\max}(K\epsilon_0^2).$$

By multiplying both LHS and RHS of this equation by $\boldsymbol{Q}$, we obtain that

$$\check{\boldsymbol{Q}} = \boldsymbol{Q} + f_{\boldsymbol{R}}(\boldsymbol{E}\boldsymbol{Q}^\top)\boldsymbol{Q} + \mathcal{O}_{\max}(K^{3/2}\epsilon_0^2).$$

In addition, by definition of $\check{\boldsymbol{R}}$, we have

$$\begin{aligned}\check{\boldsymbol{R}} &= (\boldsymbol{R}\boldsymbol{Q}+\boldsymbol{E})\check{\boldsymbol{Q}}^\top \\ &= (\boldsymbol{R}+\boldsymbol{E}\boldsymbol{Q}^\top)\boldsymbol{Q}\check{\boldsymbol{Q}}^\top \\ &= \boldsymbol{R} + \boldsymbol{E}\boldsymbol{Q}^\top - \boldsymbol{R}f_{\boldsymbol{R}}(\boldsymbol{E}\boldsymbol{Q}^\top) + \mathcal{O}_{\max}(\epsilon_0^2),\end{aligned}$$

which concludes the proof. □

C Proof of Theorem 3.4

Proof of Theorem 3.4. First, we use Lemma 3.3 to give a first-order perturbation expansion for the reduced Cholesky factor $\boldsymbol{N}^m$ of $\boldsymbol{A}^m = (\boldsymbol{N}+\boldsymbol{E}^m)(\boldsymbol{N}+\boldsymbol{E}^m)^\top$. Define $\boldsymbol{Q}^m \in \mathcal{O}_{K\times K}$ as an orthogonal matrix such that $\boldsymbol{N}^m = (\boldsymbol{N}+\boldsymbol{E}^m)\boldsymbol{Q}^{m\top}$, or equivalently, $(\boldsymbol{R}+\boldsymbol{E}^{1,m})\boldsymbol{Q}^{m\top}$ is a lower triangular matrix with positive diagonal elements. By Lemma 3.3, when $\|\boldsymbol{E}^{1,m}\|_{\max} \le \epsilon_0$ is sufficiently small, we have

$$\boldsymbol{Q}^m = \boldsymbol{I}_K + f_{\boldsymbol{R}}(\boldsymbol{E}^{1,m}) + \mathcal{O}_{\max}(\epsilon_0^2),$$

where $f_{\boldsymbol{R}}$ is defined in Lemma 3.3. By definition of $\boldsymbol{Q}^m$, we have

$$\boldsymbol{N}^m = (\boldsymbol{N}+\boldsymbol{E}^m)\boldsymbol{Q}^{m\top} = \boldsymbol{N} + \boldsymbol{E}^m - \boldsymbol{N}f_{\boldsymbol{R}}(\boldsymbol{E}^{1,m}) + \mathcal{O}_{\max}(\epsilon_0^2),$$

where we use the property $f_{\boldsymbol{R}}(\boldsymbol{E}^{1,m})^\top = -f_{\boldsymbol{R}}(\boldsymbol{E}^{1,m})$. Using this linear perturbation expansion, we are now able to characterize the Karcher mean $\widetilde{\boldsymbol{A}} = \widetilde{\boldsymbol{N}}\widetilde{\boldsymbol{N}}^\top$ (or $\widetilde{\boldsymbol{N}}$) of $\{\boldsymbol{A}^m\}_{m=1}^M$ (or $\{\boldsymbol{N}^m\}_{m=1}^M$) on the manifold $S^*(p,K)$ (or $\mathcal{L}^*(p,K)$). By the LRC algorithm, i.e., (2.3), we have $\widetilde{\boldsymbol{N}}$ is equal to $\frac{1}{M}\sum_{m=1}^M \boldsymbol{N}^m$ except that the diagonal elements of $\widetilde{\boldsymbol{N}}$ are given by

$$\widetilde{\boldsymbol{N}}_{ii} = (\prod_{m=1}^M \boldsymbol{N}_{ii}^m)^{1/M}, \quad \forall 1 \le i \le K.$$

However, when ϵ_0 is sufficiently small, $|\boldsymbol{N}_{ii}^m - \boldsymbol{N}_{ii}|$ is of order $\mathcal{O}(\epsilon_0)$ and thus

$$\widetilde{\boldsymbol{N}}_{ii} - \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{N}_{ii} = \mathcal{O}(\epsilon_0^2), \quad \forall 1 \le i \le K.$$

Therefore, we have

$$\begin{aligned}
\widetilde{\boldsymbol{N}} &= \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{N}^m + \mathcal{O}_{\max}(\epsilon_0^2) \\
&= \boldsymbol{N} + \frac{1}{M}\sum_{m=1}^{M} \left(\boldsymbol{E}^m - \boldsymbol{N} f_{\boldsymbol{R}}(\boldsymbol{E}^{1,m})\right) + \mathcal{O}_{\max}(\epsilon_0^2) \\
&= \boldsymbol{N} + \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{E}^m - \boldsymbol{N} f_{\boldsymbol{R}}(\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{E}^{1,m}) + \mathcal{O}_{\max}(\epsilon_0^2),
\end{aligned}$$

where the last equality follows from the linear property of $f_{\boldsymbol{R}}(\cdot)$. □

D Proof of Lemma 4.3

Proof of Lemma 4.3. Since $\boldsymbol{V}, \boldsymbol{\Lambda}$ denote the top K eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$, respectively, we have $\boldsymbol{\Sigma}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Lambda}$ and thus $\boldsymbol{N} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{Q}^* = \boldsymbol{\Sigma}\boldsymbol{V}\boldsymbol{Q}^*$. Similarly, we have $\widehat{\boldsymbol{\Sigma}}^m\widehat{\boldsymbol{V}}^m = \widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{\Lambda}}^m$. Since $\widehat{\boldsymbol{H}}^m$ and $\boldsymbol{Q}^*$ are both orthogonal matrices, we have

$$\begin{aligned}
(\boldsymbol{N} + \widehat{\boldsymbol{E}}^m)(\boldsymbol{N} + \widehat{\boldsymbol{E}}^m)^\top &= (\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{\Lambda}}^m\widehat{\boldsymbol{H}}^m\boldsymbol{Q}^*)(\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{\Lambda}}^m\widehat{\boldsymbol{H}}^m\boldsymbol{Q}^*)^\top \\
&= (\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{\Lambda}}^m)(\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{\Lambda}}^m)^\top \\
&= \widehat{\boldsymbol{A}}^m,
\end{aligned}$$

which concludes our proof. □

E Proof of Lemma 4.4

Proof of Lemma 4.4. The proof of this lemma is based on a first-order expansion of $\widehat{\boldsymbol{E}}^m$. Define $\mathcal{E}^m = \widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}$ and $\epsilon = \max_m \|\mathcal{E}^m\|_2/\Delta_K$. When $\epsilon \le 1/10$, by Lemma H.2, we have

$$\|\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m - \boldsymbol{V} - g(\mathcal{E}^m\boldsymbol{V})\|_{\mathrm{F}} \le 9\sqrt{K}\epsilon^2,$$

where

$$g : \mathbb{R}^{p\times K} \mapsto \mathbb{R}^{p\times K}, (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) \mapsto (-\boldsymbol{G}_1\boldsymbol{w}_1, \ldots, -\boldsymbol{G}_K\boldsymbol{w}_K),$$

with $\boldsymbol{G}_j = \sum_{i>K}(\lambda_i - \lambda_j)^{-1}\boldsymbol{v}_i\boldsymbol{v}_i^\top$ for $j \in [K]$ and $\lambda_i/\boldsymbol{v}_i$ being the ith eigenvalue/eigenvector of $\boldsymbol{\Sigma}$. By definition of $\widehat{\boldsymbol{E}}^m$, we have

$$\begin{aligned}
\widehat{\boldsymbol{E}}^m &= ((\boldsymbol{\Sigma} + \mathcal{E}^m)(\boldsymbol{V} + (\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m - \boldsymbol{V})) - \boldsymbol{\Sigma}\boldsymbol{V})\boldsymbol{Q}^* \\
&= \mathcal{E}^m\boldsymbol{V}\boldsymbol{Q}^* + \boldsymbol{\Sigma}(\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m - \boldsymbol{V})\boldsymbol{Q}^* + \mathcal{E}^m(\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m - \boldsymbol{V})\boldsymbol{Q}^* \\
&= \mathcal{E}^m\boldsymbol{V}\boldsymbol{Q}^* + \boldsymbol{\Sigma}g(\mathcal{E}^m\boldsymbol{V})\boldsymbol{Q}^* + \mathcal{O}_{\mathrm{F}}(\epsilon^2).
\end{aligned}$$

Since $\mathrm{vec} \circ g \circ \mathrm{vec}^{-1}$ is a linear mapping from $\mathbb{R}^{pK}$ to $\mathbb{R}^{pK}$, where $\mathrm{vec} : \mathbb{R}^{p\times K} \mapsto \mathbb{R}^{pK}$ is the vectorization mapping, the average of $\widehat{\boldsymbol{E}}^m$ can be expressed as

$$\frac{1}{M}\sum_{m=1}^{M} \widehat{\boldsymbol{E}}^m = \frac{1}{M}\sum_{m=1}^{M} \mathcal{E}^m\boldsymbol{V}\boldsymbol{Q}^* + \boldsymbol{\Sigma}g(\frac{1}{M}\sum_{m=1}^{M} \mathcal{E}^m\boldsymbol{V})\boldsymbol{Q}^* + \mathcal{O}_{\mathrm{F}}(\epsilon^2).$$

Thus, by the triangular inequality, we have

$$\|\frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{E}}^m\|_{\mathrm{F}} \le \|\frac{1}{M}\sum_{m=1}^{M}\mathcal{E}^m\boldsymbol{V}\boldsymbol{Q}^*\|_{\mathrm{F}} + \|\boldsymbol{\Sigma}g(\frac{1}{M}\sum_{m=1}^{M}\mathcal{E}^m\boldsymbol{V})\boldsymbol{Q}^*\|_{\mathrm{F}} + \mathcal{O}(\epsilon^2).$$

Since $\|\boldsymbol{Q}^*\|_2 = 1$ and $\|\boldsymbol{V}\|_{\mathrm{F}} = \sqrt{K}$, we have

$$\|\frac{1}{M}\sum_{m=1}^{M}\mathcal{E}^m\boldsymbol{V}\boldsymbol{Q}^*\|_{\mathrm{F}} \le \sqrt{K}\|\frac{1}{M}\sum_{m=1}^{M}\mathcal{E}^m\|_2.$$

In addition, since $\|\boldsymbol{\Sigma}\boldsymbol{G}_j\|_2 \le \Delta_K^{-1}\lambda_K$ for all $j \in [K]$, we have

$$\begin{aligned}\|\boldsymbol{\Sigma}g(\frac{1}{M}\sum_{m=1}^{M}\mathcal{E}^m\boldsymbol{V})\boldsymbol{Q}^*\|_{\mathrm{F}} &\le \Delta_K^{-1}\lambda_K\|\frac{1}{M}\sum_{m=1}^{M}\mathcal{E}^m\boldsymbol{V}\|_{\mathrm{F}}\\ &\le \Delta_K^{-1}\lambda_K\sqrt{K}\|\frac{1}{M}\sum_{m=1}^{M}\mathcal{E}^m\|_2.\end{aligned}$$

Thus, we have

$$\|\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{E}^m\|_{\mathrm{F}} \le C\|\frac{1}{M}\sum_{m=1}^{M}\mathcal{E}^m\|_2 + \mathcal{O}(\epsilon^2)$$

for some constant $C > 0$. □

F Proof of Lemma 4.5

Proof of Lemma 4.5. Similar to Lemma 4.4, the proof of this lemma is also based on the first-order expansion of $\widehat{\boldsymbol{E}}^m$. Define $\mathcal{E}^m = \widehat{\boldsymbol{\Sigma}}^m - \boldsymbol{\Sigma}$ and $\epsilon = \max_m \|\mathcal{E}^m\|_2/\Delta_K$. When $\epsilon \le 1/10$, by Lemma H.2, we have

$$\|\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m - \boldsymbol{V} - g(\mathcal{E}^m\boldsymbol{V})\|_{\mathrm{F}} \le 9\sqrt{K}\epsilon^2,$$

where

$$g : \mathbb{R}^{p\times K} \mapsto \mathbb{R}^{p\times K}, (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) \mapsto (-\boldsymbol{G}_1\boldsymbol{w}_1, \ldots, -\boldsymbol{G}_K\boldsymbol{w}_K),$$

with $\boldsymbol{G}_j = \sum_{i>K}(\lambda_i - \lambda_j)^{-1}\boldsymbol{v}_i\boldsymbol{v}_i^\top$ for $j \in [K]$ and $\lambda_i/\boldsymbol{v}_i$ being the ith eigenvalue/eigenvector of $\boldsymbol{\Sigma}$. By definition of $\widehat{\boldsymbol{E}}^m$, we have

$$\widehat{\boldsymbol{E}}^m = \mathcal{E}^m\boldsymbol{V}\boldsymbol{Q}^* + \boldsymbol{\Sigma}g(\mathcal{E}^m\boldsymbol{V})\boldsymbol{Q}^* + \mathcal{O}_{\mathrm{F}}(\epsilon^2).$$

By the triangular inequality and the fact that $\|\cdot\|_{\max} \le \|\cdot\|_{\mathrm{F}}$, we have

$$\begin{aligned}\|\widehat{\boldsymbol{E}}^m\|_{\max} &\le \|\mathcal{E}^m\boldsymbol{V}\boldsymbol{Q}^*\|_{\max} + \|\boldsymbol{\Sigma}g(\mathcal{E}^m\boldsymbol{V})\boldsymbol{Q}^*\|_{\max} + \mathcal{O}(\epsilon^2)\\ &\le \sqrt{K}\|\mathcal{E}^m\boldsymbol{V}\|_{\max} + \sqrt{K}\|\boldsymbol{\Sigma}g(\mathcal{E}^m\boldsymbol{V})\|_{\max} + \mathcal{O}(\epsilon^2),\end{aligned}$$

where the second inequality follows from the inequality $\|\cdot\boldsymbol{Q}^*\|_{\max} \le \sqrt{K}\|\cdot\|_{\max}$. Thus, to bound $\|\widehat{\boldsymbol{E}}^m\|_{\max}$, it suffices to bound $\|\mathcal{E}^m\boldsymbol{V}\|_{\max}$, $\|\boldsymbol{\Sigma}g(\mathcal{E}^m\boldsymbol{V})\|_{\max}$, and $\epsilon^2 = (\max_m \|\mathcal{E}^m\|_2/\Delta_K)^2$ separately. To give an upper bound on the first two terms, we will need Proposition 1 in Pilanci and Wainwright (2015), which is presented below for reader's convenience.

Proposition F.1 (Proposition 1 in Pilanci and Wainwright (2015))**.** *Let $\{\boldsymbol{z}_i\}_{i=1}^n \subset \mathbb{R}^p$ be i.i.d. samples generated from a zero-mean sub-Gaussian distribution with $\mathrm{Cov}(\boldsymbol{z}_i) = \boldsymbol{I}_p$. Then there exist some universal constants $C_1, C_2 > 0$ such that for any subset $\mathcal{Y} \subset \mathbb{S}^{p-1}$, we have with probability at least $1 - e^{-C_2 n\delta^2}$,*

$$\sup_{\boldsymbol{\eta}\in\mathcal{Y}}\left|\boldsymbol{\eta}^\top\left(\frac{\boldsymbol{Z}^\top\boldsymbol{Z}}{n} - \boldsymbol{I}_p\right)\boldsymbol{\eta}\right| \le C_1\frac{\mathbb{W}(\mathcal{Y})}{\sqrt{n}} + \delta,$$

where $\boldsymbol{Z}^\top = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) \in \mathbb{R}^{p\times n}$ *and* $\mathbb{W}(\mathcal{Y})$ *is the Gaussian width of the subset* $\mathcal{Y}$*. Specifically,* $\mathbb{W}(\mathcal{Y})$ *is defined by*

$$\mathbb{W}(\mathcal{Y}) = \mathbb{E}[\sup_{\boldsymbol{\eta}\in\mathcal{Y}} |\langle \boldsymbol{h}, \boldsymbol{\eta}\rangle|],$$

where the expectation is taken on $\boldsymbol{h} \in \mathbb{R}^p$*, which is a standard normal random vector.*

1: bound $\|\mathcal{E}^m\boldsymbol{V}\|_{\max}$. Denote by $\boldsymbol{e}_l \in \mathbb{R}^p$ the basis vector with value 1 at the lth entry and 0 at other entries. Then the max norm has the following expression,

$$\|\mathcal{E}^m\boldsymbol{V}\|_{\max} = \max_{l\in[p],k\in[K]} |\boldsymbol{e}_l^\top \mathcal{E}^m \boldsymbol{v}_k|.$$

Let $\boldsymbol{z}_i^m = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i^m$ and $\boldsymbol{Z}^m = (\boldsymbol{z}_1^m, \ldots, \boldsymbol{z}_n^m)^\top$. Since $\{\boldsymbol{x}_i^m\}_{i=1}^n$ are i.i.d. sub-Gaussian with mean $\boldsymbol{0}$ and covariance $\boldsymbol{\Sigma}$, $\{\boldsymbol{z}_i^m\}_{i=1}^n$ are i.i.d. sub-Gaussian with mean $\boldsymbol{0}$ and covariance $\boldsymbol{I}_p$. By definition of $\mathcal{E}^m$, we have

$$|\boldsymbol{e}_l^\top \mathcal{E}^m \boldsymbol{v}_k| = |(\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l)^\top (\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p)\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k|.$$

By the polarization equality, we have

$$\begin{aligned}(\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l)^\top (\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p)\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k = \frac{1}{2}\Big\{&(\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l + \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)^\top (\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p)(\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l + \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)\\ &- (\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l)^\top (\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p)(\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l)\\ &- (\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)^\top (\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p)(\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)\Big\}.\end{aligned}$$

Then by the triangular inequality, we have

$$\begin{aligned}\|\mathcal{E}^m\boldsymbol{V}\|_{\max} \le \max_{l\in[p],k\in[K]} \frac{1}{2}\Big\{&|(\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l + \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)^\top (\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p)(\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l + \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)|\\ &+ |(\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l)^\top (\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p)(\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l)|\\ &+ |(\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)^\top (\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p)(\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)|\Big\}.\end{aligned}$$

Since $\|\boldsymbol{e}_l\|_2 = \|\boldsymbol{v}_k\|_2 = 1$, we have

$$\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{e}_l + \boldsymbol{v}_k)\|_2 \le 2\|\boldsymbol{\Sigma}^{1/2}\|_2 = 2\|\boldsymbol{\Sigma}\|_2^{1/2},\ \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l\|_2 \le \|\boldsymbol{\Sigma}\|_2^{1/2},\ \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k\|_2 \le \|\boldsymbol{\Sigma}\|_2^{1/2},$$

for all $l \in [p]$ and $k \in [K]$. Define $\mathcal{Y}_1 \subset \mathbb{S}^{p-1}$ as follows,

$$\mathcal{Y}_1 = \left\{\frac{\boldsymbol{\Sigma}^{1/2}(\boldsymbol{e}_l + \boldsymbol{v}_k)}{\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{e}_l + \boldsymbol{v}_k)\|_2}\right\}_{l\in[p],k\in[K]} \cup \left\{\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_l\|_2}\right\}_{l\in[p]} \cup \left\{\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k\|_2}\right\}_{k\in[K]},$$

then we have

$$\|\mathcal{E}^m\boldsymbol{V}\|_{\max} \le C_1 \cdot \sup_{\boldsymbol{\eta}\in\mathcal{Y}_1} \left|\boldsymbol{\eta}^\top \left(\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p\right)\boldsymbol{\eta}\right|,$$

where $C_1 > 0$ is a constant dependent on $\|\boldsymbol{\Sigma}\|_2$. We remark here that in the proof notations C, C_1, C_2 represent some universal constants, which may vary according to the context. By Proposition F.1, there exist some universal constants $C_1, C_2 > 0$ such that the following inequality holds

$$\|\mathcal{E}^m\boldsymbol{V}\|_{\max} \le C_1\frac{\mathbb{W}(\mathcal{Y}_1)}{\sqrt{n}} + \delta,$$

with probability at least $1 - e^{-C_2 n\delta^2}$. Since $\mathcal{Y}_1$ is a finite set with cardinality $|\mathcal{Y}_1| \le Cp$ for some constant $C > 0$, by the maximal inequality (Mohri et al., 2018), the following inquality

$$\mathbb{W}(\mathcal{Y}_1) \le C_1\sqrt{\log(p)}$$

holds for some constant $C_1 > 0$. Thus with probability at least $1 - e^{-C_2 n\delta^2}$, we have

$$\|\mathcal{E}^m \boldsymbol{V}\|_{\max} \le C_1 \sqrt{\frac{\log(p)}{n}} + \delta.$$

2: bound $\|\boldsymbol{\Sigma} g(\mathcal{E}^m \boldsymbol{V})\|_{\max}$. The proof of this step is similar to that of step one. By definition of g, we have

$$g(\mathcal{E}^m \boldsymbol{V}) = (-\boldsymbol{G}_1 \mathcal{E}^m \boldsymbol{v}_1, \ldots, -\boldsymbol{G}_K \mathcal{E}^m \boldsymbol{v}_K).$$

Then we may write the max norm as

$$\|\boldsymbol{\Sigma} g(\mathcal{E}^m \boldsymbol{V})\|_{\max} = \max_{l\in[p],k\in[K]} |\boldsymbol{e}_l^\top \boldsymbol{\Sigma} \boldsymbol{G}_k \mathcal{E}^m \boldsymbol{v}_k| = \max_{l\in[p],k\in[K]} |\boldsymbol{e}_l^\top \boldsymbol{\Sigma} \boldsymbol{G}_k \boldsymbol{\Sigma}^{1/2} (\frac{\boldsymbol{Z}^{m\top} \boldsymbol{Z}^m}{n} - \boldsymbol{I}_p) \boldsymbol{\Sigma}^{1/2} \boldsymbol{v}_k|.$$

Similar to step one, by the polarization equality, the triangular inequality, and the fact $\boldsymbol{G}_k = \boldsymbol{G}_k^\top$, we have

$$\begin{aligned}\|g(\mathcal{E}^m \boldsymbol{V})\|_{\max} \le \max_{l\in[p],k\in[K]} \frac{1}{2}\Big\{ & |(\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\boldsymbol{e}_l + \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)^\top (\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p)(\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\boldsymbol{e}_l + \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)| \\ & + |(\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\boldsymbol{e}_l)^\top (\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p)(\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\boldsymbol{e}_l)| \\ & + |(\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)^\top (\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p)(\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k)|\Big\}.\end{aligned}$$

By definition of $\boldsymbol{G}_k$, we have

$$\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma} = \sum_{i>K} (\lambda_i - \lambda_k)^{-1} \lambda_i^{3/2} \boldsymbol{v}_i \boldsymbol{v}_i^\top,$$

and thus $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\|_2 \le \lambda_K^{3/2}\Delta_K^{-1}$. Since $\|\boldsymbol{e}_l\|_2 = \|\boldsymbol{v}_k\|_2 = 1$, we have

$$\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\boldsymbol{e}_l + \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k\|_2 \le \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\|_2 + \|\boldsymbol{\Sigma}^{1/2}\|_2 \le \lambda_K^{3/2}\Delta_K^{-1} + \lambda_1^{1/2}, \quad \|\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\boldsymbol{e}_l\|_2 \le \lambda_K^{3/2}\Delta_K^{-1}.$$

Define the following set $\mathcal{Y}_2 \subset \mathbb{S}^{p-1}$,

$$\mathcal{Y}_2 = \Big\{\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\boldsymbol{e}_l + \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\boldsymbol{e}_l + \boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k\|_2}\Big\}_{l\in[p],k\in[K]} \cup \Big\{\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\boldsymbol{e}_l}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{G}_k\boldsymbol{\Sigma}\boldsymbol{e}_l\|_2}\Big\}_{l\in[p],k\in[K]} \cup \Big\{\frac{\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k}{\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{v}_k\|_2}\Big\}_{k\in[K]}.$$

Then we have

$$\|\boldsymbol{\Sigma} g(\mathcal{E}^m \boldsymbol{V})\|_{\max} \le C_1 \cdot \sup_{\boldsymbol{\eta}\in\mathcal{Y}_2} \left|\boldsymbol{\eta}^\top \left(\frac{\boldsymbol{Z}^{m\top}\boldsymbol{Z}^m}{n} - \boldsymbol{I}_p\right)\boldsymbol{\eta}\right|$$

for some constant $C_1 > 0$ dependent on $\|\boldsymbol{\Sigma}\|_2$ and Δ_K. Again by Proposition F.1, we have with probability at least $1 - e^{-C_2 n\delta^2}$ that

$$\|\boldsymbol{\Sigma} g(\mathcal{E}^m \boldsymbol{V})\|_{\max} \le C_1 \frac{\mathbb{W}(\mathcal{Y}_2)}{\sqrt{n}} + \delta$$

for some universal constants $C_1, C_2 > 0$. Since $\mathcal{Y}_2$ is a finite set with cardinality $|\mathcal{Y}_2| \le Cp$ for some constant $C > 0$, by the maximal inequality (Mohri et al., 2018), we have

$$\mathbb{W}(\mathcal{Y}_2) \le C_1 \sqrt{\log(p)}$$

for some constant $C_1 > 0$. Thus with probability at least $1 - e^{-C_2 n\delta^2}$, we have

$$\|\boldsymbol{\Sigma} g(\mathcal{E}^m \boldsymbol{V})\|_{\max} \le C_1 \sqrt{\frac{\log(p)}{n}} + \delta.$$

3: bound ϵ^2. We will use the tail bound of $\|\mathcal{E}^m\|_2$ in Lemma H.1 to bound ϵ^2. By Lemma H.1, we have with probability at least $1 - e^{-C\sqrt{\frac{\delta n}{r}}}$ that

$$\|\mathcal{E}^m\|_2^2 \leq \delta$$

for some constant $C > 0$ and $r = \mathrm{Tr}(\mathbf{\Sigma})/\lambda_1 \leq p$. Then by union bound, we have with probability at least $1 - Me^{-C\sqrt{\frac{\delta n}{r}}}$ that

$$\epsilon^2 = \max_m \|\mathcal{E}^m\|_2^2/\Delta_K^2 \leq \delta,$$

for some constant $C > 0$.

Last: bound $\max_m \|\widehat{\boldsymbol{E}}^m\|_{\max}$. We will combine the results from 1 to 3 and apply a union bound to obtain the upper bound on $\max_m \|\widehat{\boldsymbol{E}}^m\|_{\max}$. In specific, by union bound, we have with probability at least $1 - 2Me^{-C_1 n\delta_1^2} - Me^{-C_2\sqrt{\delta_2 n/r}}$ that

$$\max_m \|\widehat{\boldsymbol{E}}^m\|_{\max} \leq C_3\sqrt{\frac{\log(p)}{n}} + \delta_1 + \delta_2,$$

for some constants $C_1, C_2, C_3 > 0$. Take $\delta_1 = \sqrt{\frac{\log(2Mp)}{C_1 n}}$ and $\delta_2 = \frac{\log(Mp)^2 r}{C_2^2 n}$, then we have with probability at least $1 - 2p^{-1}$ that,

$$\max_m \|\widehat{\boldsymbol{E}}^m\|_{\max} \leq C_1\sqrt{\frac{\log(pM)}{n}} + C_2\frac{\log^2(pM)r}{n},$$

for some constants $C_1, C_2 > 0$. When $n \gtrsim \log^3(pM)r^2$, with probability at least $1 - 2p^{-1}$ the following bound

$$\max_m \|\widehat{\boldsymbol{E}}^m\|_{\max} \leq C\sqrt{\frac{\log(pM)}{n}}$$

holds for some constant $C > 0$. □

G Proof of Theorem 4.6

Proof of Theorem 4.6. We will combine the results in Theorem 3.4, Lemma 4.3, Lemma 4.4, and Lemma 4.5 to prove this theorem. Recall that $\widehat{\boldsymbol{E}}^m = \widehat{\mathbf{\Sigma}}^m\widehat{\boldsymbol{V}}^m\widehat{\boldsymbol{H}}^m\boldsymbol{Q}^* - \mathbf{\Sigma}\boldsymbol{V}\boldsymbol{Q}^*$, $\mathcal{E}^m = \widehat{\mathbf{\Sigma}}^m - \mathbf{\Sigma}$, $\epsilon = \max_m \|\mathcal{E}^m\|_2/\Delta_K$, and $\epsilon_0 = \max_m \|\widehat{\boldsymbol{E}}^m\|_{\max}$. In addition, we partition $\boldsymbol{N} = (\boldsymbol{R}^\top\ \boldsymbol{B}^\top)^\top$ and $\widehat{\boldsymbol{E}}^m = (\widehat{\boldsymbol{E}}^{1,m^\top}\ \widehat{\boldsymbol{E}}^{2,m^\top})^\top$ such that $\boldsymbol{R}, \widehat{\boldsymbol{E}}^{1,m} \in \mathbb{R}^{K\times K}$ and $\boldsymbol{B}, \widehat{\boldsymbol{E}}^{2,m} \in \mathbb{R}^{(p-K)\times K}$. By Lemma 4.3, we have $\widehat{\boldsymbol{A}}^m = (\boldsymbol{N} + \widehat{\boldsymbol{E}}^m)(\boldsymbol{N} + \widehat{\boldsymbol{E}}^m)^\top$ for all $m \in [M]$, where $\boldsymbol{N}$ is the reduced Cholesky factor of $\boldsymbol{A} = \boldsymbol{V}\mathbf{\Lambda}^2\boldsymbol{V}^\top$. By Lemma 4.5, $\epsilon_0 = \max_m \|\widehat{\boldsymbol{E}}^m\|_{\max}$ is sufficiently small with high probability when n is sufficiently large. Thus, we can apply Theorem 3.4 to the LRC-dPCA algorithm to obtain the desired result. In specific, by Theorem 3.4, we have

$$\widetilde{\boldsymbol{N}} = \boldsymbol{N} + \frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{E}}^m - \boldsymbol{N}f_{\boldsymbol{R}}(\frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{E}}^{1,m}) + \mathcal{O}_{\max}(\epsilon_0^2),$$

where $f_{\boldsymbol{R}}$ is defined in Lemma 3.3. By the triangular inequality, we have

$$\begin{aligned}\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_{\mathrm{F}} &\leq \|\frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{E}}^m\|_{\mathrm{F}} + \|\boldsymbol{N}f_{\boldsymbol{R}}(\frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{E}}^{1,m})\|_{\mathrm{F}} + \mathcal{O}(\sqrt{pK}\epsilon_0^2)\\ &\leq \|\frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{E}}^m\|_{\mathrm{F}} + \sqrt{2}\|\boldsymbol{N}\|_2\|\boldsymbol{R}^{-1}\|_2\|\frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{E}}^{1,m}\|_{\mathrm{F}} + \mathcal{O}(\sqrt{pK}\epsilon_0^2),\end{aligned}$$

where the second inequality is due to the property $\|f_{\boldsymbol{R}}(\cdot)\|_{\mathrm{F}} \leq \sqrt{2}\|\boldsymbol{R}^{-1}\|_2\|\cdot\|_{\mathrm{F}}$. By Lemma 4.4, we have

$$\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_{\mathrm{F}} \leq C\|\frac{1}{M}\sum_{m=1}^M \mathcal{E}^m\|_2 + \mathcal{O}(\epsilon^2) + \mathcal{O}(\sqrt{p}\epsilon_0^2), \tag{G.1}$$

for some constant $C > 0$.

Next, we give a high probability bound on $\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_{\mathrm{F}}$ in three steps. First, by Lemma H.1, we have with probability at least $1 - e^{-\frac{\delta_1}{C_1\lambda_1\sqrt{r/(Mn)}}}$ that

$$\|\frac{1}{M}\sum_{m=1}^{M}\mathcal{E}^m\|_2 \le \delta_1,$$

for some constant $C_1 > 0$ and $r = \mathrm{Tr}(\boldsymbol{\Sigma})/\lambda_1(\boldsymbol{\Sigma})$. Second, by Lemma H.1 and the union bound, we have with probability at least $1 - Me^{-\frac{\delta_2}{C_2\lambda_1\sqrt{r/n}}}$ that

$$\epsilon \le \delta_2/\Delta_K,$$

for some constant $C_2 > 0$. Third, by Lemma 4.5, we have with probability at least $1 - 2Me^{-C_3 n\delta_3^2} - Me^{-C_4\sqrt{\delta_4 n/r}}$ that

$$\epsilon_0 = \max_m \|\widehat{\boldsymbol{E}}^m\|_{\max} \le C_5\sqrt{\frac{\log(p)}{n}} + \delta_3 + \delta_4,$$

for some constants $C_3, C_4, C_5 > 0$. By the union bound, we combine these three high probability bounds with (G.1) to obtain the desired result. In specific, with probability at least $1 - e^{-\frac{\delta_1}{C_1\lambda_1\sqrt{r/(Mn)}}} - Me^{-\frac{\delta_2}{C_2\lambda_1\sqrt{r/n}}} - 2Me^{-C_3 n\delta_3^2} - Me^{-C_4\sqrt{\delta_4 n/r}}$, the following inequality

$$\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_{\mathrm{F}} \le \mathcal{O}(\delta_1) + \mathcal{O}(\delta_2^2) + \mathcal{O}(\frac{\sqrt{p}\log(p)}{n}) + \mathcal{O}(\sqrt{p}\delta_3^2) + \mathcal{O}(\sqrt{p}\delta_4^2),$$

holds for some constants $C_1, C_2, C_3, C_4 > 0$. Take $\delta_1 = C_1\lambda_1\log(p)\sqrt{r/(Mn)}$, $\delta_2 = C_2\lambda_1\log(pM)\sqrt{r/n}$, $\delta_3 = \sqrt{\frac{\log(2pM)}{C_3 n}}$, and $\delta_4 = \frac{\log^2(pM)r}{C_4^2 n}$, then we have with probability at least $1 - 4p^{-1}$ that

$$\begin{aligned}\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_{\mathrm{F}} &\le \mathcal{O}(\frac{\log(p)\sqrt{r}}{\sqrt{Mn}}) + \mathcal{O}(\frac{\log^2(pM)r}{n}) + \mathcal{O}(\frac{\sqrt{p}\log(pM)}{n}) + \mathcal{O}(\frac{\sqrt{p}\log^4(pM)r^2}{n^2}) \\ &\le \mathcal{O}(\frac{\log(p)\sqrt{r}}{\sqrt{Mn}}) + \mathcal{O}(\frac{(\log^2(pM)r) \vee (\log(pM)\sqrt{p})}{n}) + \mathcal{O}(\frac{\sqrt{p}\log^4(pM)r^2}{n^2}).\end{aligned}$$

When $n \gtrsim (\log^2(pM)\sqrt{p}r) \vee (\log^3(pM)r^2)$, we have with probability at least $1 - 4p^{-1}$ that

$$\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_{\mathrm{F}} \le \mathcal{O}(\frac{\log(p)\sqrt{r}}{\sqrt{Mn}}) + \mathcal{O}(\frac{(\log^2(pM)r) \vee (\log(pM)\sqrt{p})}{n}).$$

In addition, when $n \gtrsim \frac{M((\log^4(pM)r^2)\vee(\log^2(pM)p))}{\log^2(p)r}$, we have with probability at least $1 - 4p^{-1}$ that

$$\|\widetilde{\boldsymbol{N}} - \boldsymbol{N}\|_{\mathrm{F}} \le \mathcal{O}(\frac{\log(p)\sqrt{r}}{\sqrt{Mn}}),$$

which concludes the proof. □

H Auxiliary Lemmas

The following lemma gives a tail bound of $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2$ in the sub-Gaussian case.

Lemma H.1 (Lemma 3 in Fan et al. (2019))**.** *Suppose $\{\boldsymbol{x}_i\}_{i=1}^n \subset \mathbb{R}^p$ are i.i.d. sub-Gaussian with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$. Let $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top$ be the sample covariance matrix, $\{\lambda_j\}_{j=1}^p$ be the eigenvalues of $\boldsymbol{\Sigma}$ sorted in descending order, and $r = \mathrm{Tr}(\boldsymbol{\Sigma})/\lambda_1$. There exist constants $C_1 \ge 1$ and $C_2 \ge 0$ such that when $n \ge r$, we have*

$$\mathbb{P}(\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2 \ge s) \le \exp\left(-\frac{s}{C_1\lambda_1\sqrt{r/n}}\right), \forall s \ge 0,$$

and $\|\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2\|_{\psi_1} \le C_2\lambda_1\sqrt{r/n}$.

The following lemma provides a first-order expansion of $\widehat{\boldsymbol{V}}\widehat{\boldsymbol{H}}$ around $\boldsymbol{V}$, where $\widehat{\boldsymbol{H}} = \operatorname{argmin}_{\boldsymbol{O}\in\mathcal{O}_{K\times K}} \|\widehat{\boldsymbol{V}}\boldsymbol{O} - \boldsymbol{V}\|_{\mathrm{F}}$.

Lemma H.2 (Lemma 8 in Fan et al. (2019)). *Let $\boldsymbol{\Sigma}, \widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{p\times p}$ be symmetric matrices with eigenvalues $\{\lambda_i\}_{i=1}^p$ and $\{\widehat{\lambda}_i\}_{i=1}^p$ (in descending order) and eigenvectors $\{\boldsymbol{v}_j\}_{j=1}^p, \{\widehat{\boldsymbol{v}}_j\}_{j=1}^p$ such that $\boldsymbol{\Sigma}\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j$ and $\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{v}}_j = \widehat{\lambda}_j\widehat{\boldsymbol{v}}_j$ for $j \in [p]$. Define $\mathcal{E} = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$, $S = \{s+1, \ldots, s+K\}$ for some fixed $s \in \{0, 1, \ldots, p-K\}$, $\boldsymbol{G}_j = \sum_{i\notin S}(\lambda_i - \lambda_j)^{-1}\boldsymbol{v}_i\boldsymbol{v}_i^\top$ for $j \in [K]$, and*

$$g : \mathbb{R}^{p\times K} \mapsto \mathbb{R}^{p\times K}, (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) \mapsto (-\boldsymbol{G}_1\boldsymbol{w}_1, \ldots, -\boldsymbol{G}_K\boldsymbol{w}_K).$$

Let $\boldsymbol{V} = (\boldsymbol{v}_{s+1}, \ldots, \boldsymbol{v}_{s+K})$, $\widehat{\boldsymbol{V}} = (\widehat{\boldsymbol{v}}_{s+1}, \ldots, \widehat{\boldsymbol{v}}_{s+K})$, $\boldsymbol{H} = \widehat{\boldsymbol{V}}^\top\boldsymbol{V}$, and $\widehat{\boldsymbol{H}} = \operatorname{sgn}(\boldsymbol{H}) \coloneqq \boldsymbol{U}_1\boldsymbol{U}_2^\top$, where $\boldsymbol{H} = \boldsymbol{U}_1\boldsymbol{\Gamma}\boldsymbol{U}_2^\top$ is the unique singular value decomposition of $\boldsymbol{H}$. If $\Delta = \min\{\lambda_s - \lambda_{s+1}, \lambda_{s+K} - \lambda_{s+K+1}\} > 0$ and $\epsilon = \|\mathcal{E}\|_2/\Delta \le 1/10$, where $\lambda_0 = \infty$ and $\lambda_{p+1} = -\infty$, we have

$$\|\widehat{\boldsymbol{V}}\widehat{\boldsymbol{H}} - \boldsymbol{V} - g(\mathcal{E}\boldsymbol{V})\|_F \le 9\epsilon\|g(\mathcal{E}\boldsymbol{V})\|_F.$$

It is worth noting that since $\|g(\cdot)\|_{\mathrm{F}} \le \Delta^{-1}\|\cdot\|_{\mathrm{F}}$, the Frobenius norm of the remainder term $\widehat{\boldsymbol{V}}\widehat{\boldsymbol{H}} - \boldsymbol{V} - g(\mathcal{E}\boldsymbol{V})$ is of order

$$9\epsilon\|g(\mathcal{E}\boldsymbol{V})\|_{\mathrm{F}} \le 9\epsilon\Delta^{-1}\|\mathcal{E}\boldsymbol{V}\|_{\mathrm{F}} \le 9\sqrt{K}\epsilon\Delta^{-1}\|\mathcal{E}\boldsymbol{V}\|_2 \le 9\sqrt{K}\epsilon^2.$$